Published in final edited form as:
*Cogn Psychol*. 2014 December ; 0: 1–27. doi:10.1016/j.cogpsych.2014.07.003.

# Word Categorization From Distributional Information: Frames Confer More Than the Sum of Their (Bigram) Parts

**Toben H. Mintz**[a,b], **Felix Hao Wang**[a], and **Vivian Jia Li**[a]

[a]Department of Psychology, 3620 McClintock Ave., University of Southern California, Los Angeles, CA 90089-1061

[b]Department of Linguistics, University of Southern California

## Abstract

Grammatical categories, such as noun and verb, are the building blocks of syntactic structure and the components that govern the grammatical patterns of language. However, in many languages words are not explicitly marked with their category information, hence a critical part of acquiring a language is categorizing the words. Computational analyses of child-directed speech have shown that distributional information—information about how words pattern with one another in sentences—could be a useful source of initial category information. Yet questions remain as to whether learners use this kind of information, and if so, what kinds of distributional patterns facilitate categorization. In this paper we investigated how adults exposed to an artificial language use distributional information to categorize words. We compared training situations in which target words occurred in frames (i.e., surrounded by two words that frequently co-occur) against situations in which target words occurred in simpler bigram contexts (where an immediately adjacent word provides the context for categorization). We found that learners categorized words together when they occurred in similar frame contexts, but not when they occurred in similar bigram contexts. These findings are particularly relevant because they accord with computational investigations showing that frame contexts provide accurate category information cross-linguistically. We discuss these findings in the context of prior research on distribution-based categorization and the broader implications for the role of distributional categorization in language acquisition.

### Keywords

grammatical category; artificial grammar learning; language acquisition; syntax; categorization



Corresponding Author: Toben H. Mintz, tmintz@usc.edu, Department of Psychology, 3620 McClintock Ave., University of Southern California, Los Angeles, CA 90089-1061, 213-740-2253.

## 1. Introduction

Grammatical categories—e.g., noun, verb, adjective, etc.—are the building blocks that structure human languages and the units over which syntactic and morphological processes operate. Across typologically diverse languages, categories govern the ordering of words in sentences and the combinations of affixes and word stems. For example, in the sentence *Anna is kicking the ball*, the stem *kick-* occurs with the inflection *–ing* and follows the auxiliary verb *is* by virtue of being a verb, and in particular, a verb marked with present progressive tense and aspect. Because syntactic processes apply to categories, the present progressive morphosyntax does not need to be learned or represented, item by item, for each verb of English; rather, a word's status as a verb is sufficient for licensing its occurrence in this construction. Moreover, when hearing an unfamiliar word in this construction—e.g., *is lorping*—an English speaker can identify *lorp* as a verb stem, and then, by virtue of the category, knows a host of other operations and constructions available to the stem. For example, the speaker knows that the stem can be inflected with a past-tense morpheme to form *lorped*, and that it can be sequenced with a different auxiliary verb to form *will lorp*, and so on. In sum, categories provide language users with an efficient and powerful means of representing regularities in linguistic structures and processes.

However, words do not appear in utterances explicitly marked with category information, so there is a puzzle as to how speakers first acquire the knowledge that allows them to leverage the power of categories. Consider the previous *lorp* example: An English speaker can use her implicit knowledge of English syntax to identify the structural position of *lorp* in the sentence as the head of a verb phrase (VP), and thus categorize *lorp* as a verb. But novice English learners cannot call on this knowledge, as it is precisely the mapping of surface strings—e.g., *is lorping*—to syntactic structures—e.g., VP—that they are in process of learning. Learners must therefore be able to assign at least some words in an utterance to categories in order to be able to learn about the abstract grammatical patterns in their language in the first place. How do language learners take the first steps in assigning words to categories? What information could they use to initially categorize words?

In this paper, we provide new evidence that learners perform *distributional analyses* of the sequences of words in their input and form categories of words that appear in similar distributional contexts. For example, using distributional information, a learner could analyze the utterance *Can you lorp it?* and categorize *lorp* with other verbs, not by identifying the position of *lorp* as the head of a VP, but by categorizing it with other words that occur in similar contexts. This is because words in English that are immediately surrounded by *you* and *it* are almost exclusively verbs. While other studies have presented similar evidence (Mintz, 2002; Reeder, Newport, & Aslin, 2013), our findings are significant because they demonstrate that learners are especially responsive to a particular type of distributional pattern called a *frequent frame* (like the English *you_it* frame just mentioned), which has been shown computationally to be an especially accurate source of grammatical category information cross-linguistically (Chemla, Mintz, Bernal, & Christophe, 2009; Erkelens, 2009; Stumper, Bannard, Lieven, & Tomasello, 2011; Wang, Hohle, Ketrez, Kuntay, & Mintz, 2011; Weisleder & Waxman, 2010). Our study thus sheds light on the particular kinds of distributional patterns to which human learners attend, and

their potential relevance in human language acquisition. Furthermore, some studies have only found evidence of distributionally-based categorization in situations where grammatical categories were also marked by converging sources of information, such as semantic or phonological information (Braine, 1987; Frigo & McDonald, 1998; Gomez & Lakusta, 2004; Smith, 1966). Counter to those findings, we show that learners can acquire categories solely from distributional information (Mintz, 2002; 2011; Reeder et al., 2013), and we propose an account that unifies the apparent discrepancies in prior research on distributional categorization.

### 1.1. Grammatical Categories and Distributional Analyses

The idea of associating lexical categories with distributional patterns can be traced at least to the beginning of modern linguistic theory. Indeed, in early accounts, categories were defined by co-occurrence patterns (Bloomfield, 1933), and some theories treated category labels as mere notational conveniences to stand in for the distributional patterns in which words occurred (Harris, 1951). In language acquisition, the proposal that learners initially categorize words using distributional information plays a role in diverse approaches (Gomez & Gerken, 1999; Gomez & Lakusta, 2004; Maratsos, 1982; Maratsos & Chalkley, 1980; Mintz, 2002; 2003; Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998; Reeder et al., 2013; St. Clair, Monaghan, & Christiansen, 2010; Tomasello, 2000; Wang et al., 2011 among others). The central idea is that the abstract syntactic structures that involve lexical categories—e.g., syntactic phrases—give rise to regularities and patterns in the actual sequences of words (or morphemes) in utterances, that is, in the sequences to which the learner is exposed. The hypothesis is that these patterns are sufficient for making an initial assignment of some words to categories. To take a straightforward example, while a syntactic description of a noun phrase in English is abstract (say, [$_{NP}$ DET(erminer) Noun-NUM]), it nevertheless results in actual phrases that have a discernable pattern: ***the*** *car*, ***the*** *road*, ***the*** *story*, ***the*** *car****s***, ***the*** *road****s***, ***the*** *stori****es***, ***a*** *car*, ***a*** *road*, ***a*** *story*, ***some*** *car****s***, ***some*** *road****s***, ***some*** *stori****es***, and so on. The nouns in these phrases share a range of overlapping lexical (and morphological) contexts. If learners could detect these patterns, they could, in principle, categorize words together by virtue of their occurrence in overlapping sets of patterns.

The idea of distributional analysis is an important component even of theories that posit different means for initially grouping words together. For instance, Pinker's *semantic bootstrapping* theory (Pinker, 1984) incorporates "structure dependent distributional learning." This is essentially a form of distributional analysis that is constrained to operate over syntactic constituents, such as NPs and VPs. For example, a learner might carry out a distributional analysis of a noun phrase to discover that a previously uncategorized word is a determiner. A critical aspect of structure dependent distributional learning is that the learner must already have acquired sufficient syntactic knowledge to be able to identify syntactic constituents. It was therefore not intended as a mechanism for initially discovering category relationships among words. Indeed, Pinker argued that the distributional patterns that children encounter are too variable to be useful as an initial source of category information, and would even be misleading (for discussion, see Mintz et al., 2002; Redington et al., 1998; Reeder et al., 2013).

The recent interest in the role of distribution-based categorization in language acquisition has been driven in part by research analyzing distributional patterns in child-directed speech across a variety of languages. These analyses challenged criticisms of distributional approaches by showing that computationally simple distributional patterns in fact provide very accurate categorizations of words and morphemes (Cartwright & Brent, 1997; Chemla et al., 2009; Mintz, 2003; Mintz et al., 2002; Redington et al., 1998; St. Clair et al., 2010; Weisleder & Waxman, 2010). One discovery, out of which the present study was developed, was of a particularly powerful yet simple distributional pattern called a *frequent frame* (Mintz, 2003). Mintz defined a frequent frame as a frequently occurring two-word sequence (the frame) in which the two words were separated by one intervening word position. The frame *you_it*, discussed earlier, is one such frame in English child-directed speech, and contains a large number of different verbs (Mintz, 2003). Researchers have found similar patterns in French (Chemla et al., 2009) and Spanish (Weisleder & Waxman, 2010). That said, the category information provided by frequent lexical frames in child-directed German —a language that has more variable word order—is not as robust as in those other languages (Stumper et al., 2011; Wang et al., 2011). This result could be taken as evidence that frequent frames, and perhaps distributional information in general, may be less useful for languages with freer word order. However, analyses in which *morphemes* in German were treated as the distributional units, frequent (morpheme) frames resulted in category generalizations that approached the accuracy of those in English (Wang et al., 2011). Moreover, morphological frequent frames also yielded highly accurate categories in Turkish, a language with rich inflectional morphology and minimal syntactic regulation of word order (Wang et al., 2011; for related findings in Dutch, see Erkelens, 2009). Those findings suggest that distributional information, when adapted to the level of representation where the grammar has most influence on surface sequences, is a cross-linguistically a rich source of category information.

Although frames are relatively simple distributional contexts, other studies investigated distributional analyses based on even simpler *bigrams*, in which one word is treated as a context for categorizing an immediately adjacent word—e.g., ***the*** *car*, ***a*** *car*, ***the*** *book*, ***a*** *book*, *car* ***is***, *book* ***is***, etc. In those studies, a target word's distributional context was represented as a multidimensional vector that encoded information about the identity of words occurring in sequences with the target word, including the position relative to the target word and the frequency at that position. The context of each target word was then compared to every other one by numerically comparing their context vectors; a hierarchical cluster analysis was used to group words together that had highly overlapping distributional contexts. This procedure resulted in a number of groupings that were very accurate with respect to linguistic categories, but also resulted in some relatively large categories that were incoherent with respect to actual form-class categories (Mintz et al., 2002; Mintz & Newport, 1995; Redington et al., 1998). Thus, although bigrams are formally simpler contexts, in some analyses they resulted in less accurate categories compared to frames.

A more recent approach using a connectionist modeling framework and a supervised learning algorithm evaluated models' ability to generalize after being trained on the association of specific contexts with categories (St. Clair et al., 2010). St. Clair et al.

analyzed a variety of models that considered different types of distributional contexts, one of which they called *flexible frames*. Flexible frames are similar to frequent frames in that they involve trigram sequences in which the first and last word are contexts for categorizing the medial word. However, the first and last words of a flexible frame are not explicitly treated as a unit (i.e., as a distinct frame). St. Clair et al. compared connectionist models trained to associate flexible frames with categories to those trained to associate frequent frames with categories. In the case of flexible frames, a training trial for a trigram sequence *XTY* involved activating one input node for each of the context words in the two bigrams (i.e., one for the *X*_ context, one for the _*Y* context), and activating an output node corresponding to the category of the target word, *T*. In contrast, training in their frequent frame model involved activating one input node for the *X*_*Y* frame and the output node for the category of the target, *T*. Thus, for a given trigram, the representations of flexible frames and frequent frames captured the same two contexts words, but in the flexible frames model, they were presented simultaneously as two independent contexts. Thus, the "flexible" moniker captures the idea that the model represents the framing elements as independent words participating in bigram patterns with the target, but, by virtue of their simultaneous presentation, it can also incorporate the frame organization (i.e., the co-occurrence of X_ and _Y) to the degree that it benefits categorization. St. Claire et al. reported higher categorization scores for flexible frames over frequent frames, and argued that a learner would be more successful by considering bigrams (in a trigram context) as opposed to frames, or bigrams alone. Thus, similar to previous computational analyses, St. Clair et al. (2010) concluded that distributional patterns provide a rich source of category information, but they argued that considering the bigrams that constitute frames is advantageous over considering only frames.

In summary, the broad range of computational analyses on typologically diverse languages suggest that distributional information computed by relatively simple learning mechanisms could provide a linguistically informative initial sorting of words into categories. However, proposals differ on the types of distributional information they claim is important for human learners, in particular, whether learners attend primarily to bigrams[1] or to frame contexts. Moreover, the computational findings offer no indication of whether learners actually use distributional information to categorize words, and if so, what kind. Thus, a key question for this study was to assess learners' ability to form categories when category information is carried by frames compared to when category information is carried only by patterns involving bigrams.

### 1.2. Behavioral Investigations of Distributional Category Learning

Many of the early behavioral studies that investigated distributional analyses suggested that learners cannot use distributional information alone as a basis for categorization (Braine et al., 1990; Frigo & McDonald, 1998; Gerken, Wilson, & Lewis, 2005; Smith, 1969). Subjects in those studies were shown to generalize from a particular distributional cue only when the cue correlated with some other source of category information, for example, semantic or

[1]We use the term *bigram context* to refer to a sequence where one word provides the context for categorizing an adjacent word. Technically, the context is a *unigram*, but it is conventional to refer to the entire bigram sequence when describing this scenario.

phonological information. This issue is important because, although semantic and phonological properties do pattern with distributional information and category status, there are important limitations to how informative those sources might be for infants' initial category generalizations. In the case of semantic co-occurrences, the correspondences between semantic and syntactic categories in natural language are not nearly as tight as they typically are in laboratory settings. In some cases, semantic information is completely uncorrelated, as in the case of grammatical gender marking (Maratsos, 1982). In other cases, semantic-syntactic correspondences go against the canonical patterns. For example, while one might assume that nouns typically refer to concrete objects, many very common nouns do not: For instance, the noun in, *a walk*, refers to a very *actiony* concept. Examples like this are by no means the exception in speech to children (Maratsos & Chalkley, 1980). Moreover, there is considerable controversy in the field as to whether the semantic referents of all but a few nouns can be gleaned from situational observation, without first knowing something about their category and the syntactic structures in which they are situated (Fisher, Hall, Rakowitz, & Gleitman, 1994; L. R. Gleitman, 1990; L. R. Gleitman & Gleitman, 1992; L. R. Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). There is also evidence that access to semantic information might *hinder* the learning of distributional associations (Arnon & Ramscar, 2012). It is therefore reasonable to be cautious in generalizing from the laboratory findings regarding correlated semantic cues to natural language acquisition more broadly.

In the case of phonological and syntactic co-occurrences, it is important to consider that while there is considerable cross linguistic overlap in the cues that differentiate open and closed class words (Shi, Morgan, & Allopenna, 1998), the particular phonological properties that correlate with grammatical categories within those classes (e.g., between nouns and verbs) differ across languages, and in some instances they even pattern in contradictory ways (Monaghan, Christiansen, & Chater, 2007). Since the learner cannot know ahead of time which phonological property or properties are correlated with grammatical category for the language they are learning (or *how* they are correlated), it would be necessary for them to have a representation of at least some rudimentary categories in order to determine which phonological properties were relevant. In principle, distributional information could provide this initial categorization. For these reasons, it is worthwhile to understand learners' capacities and limitations in forming category generalizations from distributional information alone. Proto-categories formed in this way could then serve as a foundation for discovering how other cues (e.g., phonological cues) pattern with those categories.

Interestingly, many (although not all) of the experiments that failed to find evidence of categorization in the absence of correlated cues involved bigram patterns in which one word was the context for categorizing an immediately adjacent word. In contrast, more recent studies (with adults and infants) that involved distributional information that was more similar to the frames described earlier have shown that learners can use distributional information alone to categorize unfamiliar words (Mintz, 2002; 2006a; 2011; Reeder et al., 2013). For example, Mintz (2006a) exposed 12-month-old infants to nonsense words (e.g., *lonk*, *deeg*) in otherwise normal English sentences, where the words appeared in several frequently occurring noun frames (e.g., *I see the lonk in the room!*), or in several frequently

occurring verb frames (e.g., *I see you deeg it!*). The training material thus involved distributional patterns within the experiment itself (e.g., nonsense verbs occurred in many of the same frames), as well as links to distributionally informative patterns in English. Infants listening times when tested on 'grammatical' and 'ungrammatical' sentences involving the nonsense words suggested that they had categorized the words. Of course, since the words, although meaningless, were in English constructions, it is possible that 12-month-olds brought to bear syntactic knowledge that they had already acquired to categorize the nonsense words (as in the earlier *kick* example ). In that case, categorization would not have been due purely to the distributional patterns in the word sequences, but could have been aided by the learners' syntactic knowledge of English (rudimentary though it may be at 12 months).

However, there are studies with adults that clearly involve only distributional information. Mintz (2002) exposed adults to miniature artificial languages in which target words occurred within frames. Learners were exposed to training material in which a set of target words occurred in a highly overlapping set of frames, providing distributional support for treating the target words as members of a category. Some contexts occurred with only a subset of the target words during training, yet in testing, learners' responses indicated that they generalized from the contexts in which all the target words occurred to the more limited contexts, and treated the occurrences of previously unattested words in those contexts as natural. The best explanation of these results was that learners treated the target words as a category, and generalized properties that were observed for only some of the members—i.e., a specific context in which they appeared—to all members.

Reeder et al. (2013) further explored the conditions under which learners generalize from distributional information. As in Mintz (2002), they exposed learners to target words in the context of frames (although, they did not argue that frames, as such, played a critical role in the successful categorization behavior in their experiments). However, in a series of experiments they further manipulated the distributional properties of the contexts to investigate how two variables—density and overlap—contribute to category generalization. Density refers to the number of different contexts in which the target words occur, and overlap refers to the degree to which different target words share contexts (that is, how much their sets of contexts overlap). In particular, they asked how these variables influence learners' treatment of gaps in a distributional paradigm: Under what conditions do they judge the missing sentences to be grammatical (thus, concluding that the gaps were accidental), and under what conditions do the gaps signal that the missing sentence is not part of the language (i.e., ungrammatical)? Reeder et al. found that when overlap was robust, learners treated unattested but distributionally supported strings as grammatical, as though their absence during training was accidental. However, in conditions where there was less overlap in the sets of contexts in which target words occurred, learners were less likely to generalize grammaticality to sentences that they had not heard during the training phase. The findings suggest that in the latter situation, learners did not form categories, even though there was some distributional support for category generalization. Thus, as well as strengthening support of the hypothesis that learners can form categories from distributional

information alone, Reeder et al. (2013) demonstrated the importance of context overlap on distributional category generalization.

Returning to the question of the particular distributional contexts that facilitate category generalization, the range of factors that varied across behavioral studies on distributional categorization make it difficult to compare, post hoc, the effectiveness of frames and bigrams. Moreover, in cases where categorization was found in frame contexts, the benefit provided by frames *per se* is unclear because every instance in which a word occurs within a frame, it also occurs in two bigrams. Hence, it is possible that subjects' category generalization in frame constructions (Mintz, 2002; Reeder et al., 2013) was really due to the presence of multiple bigram contexts (St. Clair et al., 2010), not frames *per se*. Furthermore factors such as degree of overlap (Reeder et al., 2013) might have been more favorable for categorization and generalization in the studies just discussion, in comparison to prior studies that failed to show categorization from bigram patterns without converging cues.

### 1.3. The Present Study

In this study we addressed this question by explicitly comparing learners' category generalizations when distributional information was supplied by frames versus when it was supplied only by bigrams. We ask whether frequent frames offer some special advantage in distributional category learning, beyond simply being the sum of two bigrams. This is an important question because understanding the distributional contexts that facilitate categorization provides critical insight into the mechanisms that may be involved in a fundamental process of language acquisition.

To address this question, we carried out three experiments in which we exposed adult subjects to samples of a simple artificial language. The language had no meaning, but the distributional patterns in which certain words occurred provided a basis for treating them as members of a category. Across experiments, we varied the nature of the distributional information, in particular, whether the distributional information included frequent frames, or only bigrams. Critically, the information contained in the bigrams was similar across experiments, while the information contained by frames varied, allowing us to determine the degree to which learners depended on frames to categorize words.

## 2. Experiment 1

In Experiment 1 we provided distributional information in frequently occurring frames, and by consequence in the bigram patterns that necessarily result. The goal of this experiment was to establish whether subjects would generalize the distributional privileges of a word to contexts in which the word has not been experienced, based on its similarity of patterning with other words. Such a finding would demonstrate that learners formed a category using distributional information in the linear word sequences and generalized the word's distributional privileges from its category membership. This would replicate findings from recent work, but with different materials and implementing more complex distributional properties (see section 2.1.2). But critically, the results from Experiment 1 will be compared to those of Experiments 2 & 3 to elucidate the particular kinds of distributional properties to

which learners attend when categorizing words. In particular, the comparison will shed light on the question of whether learners use frames when categorizing words from lexical co-occurrence patterns, above and beyond the two bigrams that make them up.

## 2.1. Methods

**2.1.1. Subjects—**Twelve college students participated in the experiment for partial satisfaction of course requirements. Subject were randomly assigned to one of the two counterbalanced training groups, resulting in six subjects per group. We established an exclusion criterion such that subjects performing at or below 50% correct on trials that compared repeated to scrambled sentences would not be used, however all subjects performed above the exclusion threshold.

**2.1.2. Materials and Design—**Broadly, the language was comprised of two distributional paradigms (A and B), each of which provided distributional evidence for a distinct grammatical category. Each sentence in the language belonged to one of these paradigms and contained a critical three word sequence (trigram) in which the first and last words could be thought of as a frame, or context, and the medial word as a target word. There were also additional words added optionally before or after the critical trigram. The primary purpose of these optional words was to vary the absolute position of the critical trigram words, as well as relative position to the sentence boundaries (see Reeder et al., 2013). The optional words were the same in both paradigms (*ghen* or *dap* could occur before the trigram, and *tiv* or *nud* could occur after), and their occurrence patterns were such that neither the presence or absence of an optional word (in either position) nor the particular words themselves were predictive of either paradigm, of any other word, or of any frame. However, the trigrams were the critical sequences that carried category information, and we focus on the trigram structure in the remainder of this section.

**2.1.2.1. Training materials:** In each paradigm, the critical trigrams were constructed from three frame-initial words and three frame-final words (resulting in a total of nine unique frames), with one of six target words occurring in trigram-medial position (i.e., surrounded by the frame). The training materials were designed to provide distributional evidence that the target words within a paradigm belonged to the same category. An exhaustive paring of the nine frames with each target word would yield 54 unique trigrams per paradigm (the cells in Table 1). Exposure to the complete paradigm would provide a learner with strong distributional evidence that all the medial words within the paradigm belonged to the same category, because the contexts in which the words occur would be identical. However, we trained subjects on only a subset of the full paradigm, to allow us to test for generalization using the untrained items. The cells in Table 1 labeled ‘E1’ indicate which trigrams occurred in the training sentences in paradigm A, and Table A.1 shows the trigrams that occurred in paradigm B training sentences. Critically, the training subset contained distributional evidence that all the medial words in a paradigm *could* occur in the same contexts, even though there were some contexts in which they did not appear during training. For example, while the training material did not contain a sentence in which *lartsu* occurred in the *ghire_blit* frame, it did contain sentences in which *lartsu* occurred in many of the same frames (i.e., distributional contexts) as words that *did* occur in the *ghire_blit* frame (namely,

*channer*, *lowfa*, *refton*, and *sykteb*; Table 1). A learner could generalize from the shared contexts and conclude that *lartsu* has the same co-occurrence privileges as those other words, and thus should be expected in the *ghire_blit* frame, even though the learner did not hear it in that context.

There were 72 unique trigrams in the training set, made from 18 frames (nine per paradigm), that occurred with subsets of the six target words. The cells with an E1 in Table 1 for paradigm A and in Table A.1, in the Appendix, for paradigm B show the specific subsets of four target words that occurred with each frame. The particular pairing of frames and target words was constrained by the design of the test items, described in the next section, and their relationship to the training material. The full training set consisted of three repetitions of each of the 72 trigrams; each repetition involved a different arrangement of optional words and empty optional positions. For example, the *choon channer glaik* trigram occurred in the sentences *choon channer glaik nud*, *ghen choon channer glaik*, and *dap choon channer glaik tiv*. The total training set thus consisted of 216 sentences.

**2.1.2.2. Test Materials:** Four types of test sentences accomplished the following broad goals: *novel-grammatical* and *novel-ungrammatical* sentences assessed categorization of target words, and *repeated* and *scrambled* sentences provided a baseline for performance and a gauge of general attention to the experimental task. All test sentences were three words long.

***2.1.2.2.1. Novel-Grammatical and Novel-Ungrammatical Test Sentences:*** Novel-grammatical sentences were a subset of the 'missing' training sentences, that is, a subset of the empty cells in Table 1 and Table A.1. In particular, novel-grammatical items involved unattested frame and target word combinations. A further constraint was that the target word did not occurr with either individual test frame word during training. For example, the sequence *swetch lowfa klide* was a novel-grammatical test sentences because the training set did not contain the target word *lowfa* in the *swetch_klide* frame, nor did it include any sentence that contained the bigrams *swetch lowfa* or *lowfa klide*. Yet *lowfa* shares other contexts in the training set with target words that do occur in the *swetch_klide* frame; if learners categorize *lowfa* with the other target words based on the distributional similarity, they could judge *swetch lowfa klide* to be an acceptable sequence. Hence, the *swetch_klide* frame provides a strong test of categorization because if subjects judge *swetch lowfa klide* to be acceptable, their judgment cannot be due to encountering a familiar sequence of words. (In contrast, while *lowfa* does not occur in the *ghire_klide* frame in the training set, it does occur after *ghire*, in the *ghire_blit* frame. Hence, the *ghire_klide* frame does not provide a strong test of categorization, because learners could judge the trigram to be acceptable based on the attested *ghire lowfa* sequence. Hence, *ghire lowfa klide* did not qualify as a novel-grammatical test item.) In each paradigm, two trigrams satisfied the constraints just described, yielding four novel-grammatical trigrams across paradigms. Again, these items conformed to the category structure of the training set, but contained no subsequences that were attested in the training materials. The novel-grammatical test items for paradigm A (Table 1) were *swetch lowfa klide* and *choon pooda blit*; the novel-grammatical items for paradigm B (Table A.1) were *ghip puziv voy* and *dess mirshow sowch*.

Of course, the novel-grammatical trigrams do contain certain trained sequences, in particular, the non-adjacent sequences that form the frames (e.g., the frame *swetch_klide* occurs in the training set). Novel-grammatical strings might thus be acceptable to subjects based on the familiarity of the non-adjacent pattern (Gomez, 2002). In order to ensure that our measurements assessed categorization and not frame familiarity, we devised additional test materials that controlled for the familiarity of the non-adjacent sequences: Each novel-grammatical test item had a novel-ungrammatical counterpart that had an identical frame, but a target word from the other paradigm. For example, the novel-ungrammatical sentence *swetch puziv klide* was formed by taking the frame from the novel-grammatical test item *swetch lowfa klide* and replacing the medial word with one from paradigm B. Critically, novel-grammatical and novel-ungrammatical test items were identical with respect to their superficial similarity to the training sentences: Specifically, frames in each type of sentence were identical, and target words were matched in overall frequency as well as their frequency in the absolute position in the test sentences (i.e., the second word), and each subsequence had a transitional probability of 0. The only systematic difference between novel-grammatical and novel-ungrammatical test items was distributional: A target word in a novel-grammatical sentence belonged to the same distributional category as target words that occurred in the novel-grammatical frame during training; target words in novel-ungrammatical sentences did not.

***2.1.2.2.2. Repeated and Scrambled Test Sentences:*** There were two types of test sentence in addition to the novel-grammatical and novel-ungrammatical types. Repeated test sentences were trigrams that were exact duplicates of ones that occurred in the training set. Scrambled test sentences disrupted many sequence and position regularities that were present in the training set. Specifically, words that only occurred at edges of the critical trigrams in the training sentences occurred in the middle of scrambled sentences, the relative order of words with respect to each other was also disrupted, and words from the two paradigms were mixed. For example, the scrambled test sentence *voy blit channer* placed a trigram-final word from paradigm B in a trigram-initial position, followed by a trigram-final word from paradigm A in the trigram-medial position, followed by a target word from paradigm A in trigram-final position. We reasoned that subjects who were attending to the training material would judge repeated sentence to be familiar but scrambled sentences to be very unfamiliar and unexpected with respect to the structure of the training material. Subjects' performance on these sentence types can be treated as an upper bound on performance in this procedure. In addition, if a subject performs poorly on these sentences, it suggests that they were probably not attending to the experimental materials and therefore should not be expected to generalize.

We constructed 32 two alternative forced choice (2AFC) test trials by combing each novel-grammatical sentence with each novel-ungrammatical sentence (4×4=16) and each repeated with each scrambled sentence (4×4=16). Since test sentences must be presented sequentially within a trial, our pairing also specified which item was first, such that a given sentence was first on two of the four trials in which it was present. From the 32 2AFC trials created in this manner, we created 32 more that had the identical pairings, but switched the order of

sentences within each pair, to counterbalance the orders in the original list. Each subject was tested on both lists, as described in section 2.1.3.2.

**2.1.2.3. Counterbalanced Training Sentences:** From the original 216 training sentences (section 2.1.2.1) we created a counterbalanced training set that differed only with respect to the target words that were involved in the novel-grammatical and novel-ungrammatical test items (section 2.1.2.2). In particular, the two paradigm-A target words (*lowfa* and *pooda*) were exchanged with the counterpart paradigm-B words (*puziv* and *mirshow*) to create the counterbalanced training set. Since the same test items were associated with the two counterbalanced training sets, an item that was novel-grammatical in one training set was novel-ungrammatical in the other training set. This ensured that greater responses to novel-grammatical over novel-ungrammatical sentences could not result from idiosyncratic properties of the items themselves.

**2.1.2.4. Stimulus recording:** The auditory source materials were recordings of the words of the artificial language, spoken by a female speaker of American English. The speaker digitally recorded the words into a master computer file as they were displayed, one at a time, on a computer monitor approximately every 1.5 s. She produced each word at approximately the same rate of speech with list intonation. We then digitally segmented the master list into individual word files that began at the onset of the word with a silence pad at the end so that the each word's sound file had a duration of 1 s, including the silence pad. Each word file was then shortened to .8 s using an algorithm in Praat (Boersma & Weenink, 2009) that compressed the presentation rate without sacrificing intelligibility (e.g., pitch was not altered), to allow more training material to be presented in a fixed amount of time. We used the computer program shntool (Jordan, 2009) to automatically splice together words to create the training and test sentences. In the resulting sentences, words occurred every .8 s with variable length silence between word onsets.

Stimuli for the experiment included the auditory sentences as well as additional text versions of the training sentences. The text version of each training sentence was displayed at the same time that the auditory version was presented. The purpose of the text was simply to facilitate processing of the unfamiliar words.

**2.1.3. Apparatus and Procedure—**Subjects were tested one at a time, in a quiet room, seated at a desk. Stimulus presentation and subjects' responses were controlled by an Apple Mac Mini computer that was connected to a keyboard, a monitor, and headphones. The software package PsyScope X;B53 (Cohen, Macwhinney, Flatt, & Provost, 1993) controlled the presentation of materials and collection of keyboard response data.

In order to determine how the quantity of experience with the distributional patterns influences category generalizations, training and testing was administered in two blocks. At the beginning of the first block subjects were given the following instructions:

> "You will listen to a number of sentences in a made up language. Simply pay attention to the sentences, and respond to the quizzes presented. After you have heard the sentences repeated a number of times, you will be asked some test

> questions to test your memory for what you heard. Before you hear the sentences, you will hear a list of all the words in the made up language, so they won't be completely strange to you. You can use this time to adjust the volume on the headphones so it's at a comfortable level using the volume keys (F11 & F12) at the top of the keyboard."

The experimenter gave the subject the opportunity to ask questions, and then the subject indicated via the keyboard that he or she understood the instructions. After presentation—auditory and visual—of the list of words in the language, the following prompt was displayed on the screen:

> "In the next part of the experiment, you will listen to some sentences in the made up language. Simply pay attention to these sentences. At various points you will be quizzed on what you hear. The sentences will play for about 12 minutes. Click the mouse to begin listening to the sentences."

**2.1.3.1. Training Phase:** The experiment started with the training phase, in which the 216 sentences from the appropriate counterbalanced set were played to the subjects via the headphones and concurrently displayed on the monitor. Recall that the training set contained three different sentences for each critical trigram that differed only in the optional words (section 2.1.2.1). The computer presented the sentences in three blocks, such that each instance of a critical trigram was in a different block. The order of the sentences was randomized within blocks.

Every twelve sentences the computer stopped the presentation of the training material and displayed the following question on the monitor: "Which of the following words was in the last sentence you heard? Please type the number corresponding to the correct word below," followed by a numbered list of six words from the language and the word 'none', always as the 7$^{th}$ choice. The particular set of options was chosen by the computer at random from a predetermined set of choice lists, and there was always exactly one correct answer. These 'quizzes' were designed to encourage subjects to attend to the material.

**2.1.3.2. Test Phase:** After the training phase, the computer displayed the following prompt:

> "Next, you will listen to some pairs of sentences. Please decide whether the first sentence or the second sentence is most familiar to you. If the first sentence is more familiar, press the "1" key. If the second sentence is more familiar, press the "2" key. Click the mouse to begin the test phase."

Next, the computer presented 32 2AFC test trials (16 that paired a repeated and scrambled sentence, and 16 that paired each novel-grammatical with each novel-ungrammatical test sentence). Each sentence in a trial was played once, and the sentences were also displayed on the screen, with the first sentence presented on the left and the second on the right. Subjects could take as long as they wanted to enter a response. The order of presentation of the test trials was random.

After the test phase, the computer displayed the following prompt: "You're almost through! Next, you will hear the sentences repeated some more times. Just listen again and do the

quizzes, like before, and then you'll be tested again. Click the mouse to continue. Good luck!" The subject's mouse click initiated the second block of the experiment, which followed the same training and testing procedure as the first block, but items in each phase followed a new random order. In addition, in the test phase, the order of the sentences within each 2AFC pair was switched in the second block. Subjects were given the opportunity to take a brief break before beginning the second training block. The second training and testing blocks were administered following the same procedure as the first.

### 2.2. Results

We first briefly discuss the quiz results. The primary purpose of the quizzes was to encourage subjects to attend to the training material and to confirm that they were attending to some minimal degree. We therefore report the analyses of the quiz data for all experiments here. Since there were seven choices and one correct answer, the probability of a correct response by chance was 1/7, or approximately 5 out of 36 questions (18 in each training block). All subjects were significantly above chance in choosing the correct responses to quiz questions, across all experiments ($\chi^2 > 15$ for each subject, $p < .001$ for each subject), indicating that subjects were attending to the training material to some degree. Figure 1 graphs the proportion of quiz questions correctly answered, for each subject, organized by experiment.

Turning now to the test data, we coded the 2AFC responses as "correct" (1) or "incorrect" (0), as a function of the trial type. For trials that compared repeated sentences to scrambled sentences, we coded responses as correct if the subject chose the repeated sentence, and incorrect if the subject chose the scrambled sentence. For trials that compared novel-grammatical sentences to novel-ungrammatical sentences, we coded responses as correct if the subject chose the novel-grammatical sentence, and as incorrect if the subject chose the novel-ungrammatical sentence. Average proportion correct across subjects for novel-grammatical versus novel-ungrammatical trials was 57.8% and for repeated/scrambled trials was 72.4%. Figure 2 shows individual subject means for each comparison type.

To statistically evaluate subjects' performance against chance (50%) and to assess the influence of the experimental variables on subjects' performance, we used a logistic regression model (Jaeger, 2008) with the binary response (correct = 1, incorrect = 0) as the dependent variable. We first ran a model with intercept, counterbalancing group (1 or 2), trial type (repeated vs. scrambled, or novel-grammatical vs. ungrammatical, 2AFC), block (first or last), and paradigm (A or B) as fixed effects, controlling for random subject variance on intercept and slopes of within-subject variables (i.e., all but the counterbalancing variable). The novel ungrammatical and scrambled items contained words from both paradigms, by definition, so we labeled each trial by the paradigm (A or B) associated with the grammatical/repeated string, depending on trial type. Since there was no effect of counterbalancing group, and no interaction of that variable with any other variable (or their interactions), we ran a model without the counterbalancing variable, to increase power. The results of the analysis are summarized in Table 2. Unsurprisingly, there was a main effect of trial type, with subjects performing better on the repeated vs. scrambled 2AFC compared to the novel-grammatical vs. novel-ungrammatical. There were also main effects of block and

paradigm, as well as significant two- and three-way interactions between all variables. Figure 3 plots mean accuracy, broken down by these variables. Visually, one can see that the nature of these effects and interactions is that, overall, subjects were somewhat more accurate on the repeated vs. scrambled 2AFC, and that subjects were generally consistent across blocks, however subjects were less accurate in the first block of novel-grammatical vs. novel-ungrammatical trials in Paradigm A.

These impressions were backed up by further targeted logistic regressions. Analyzing results only for repeated vs. scrambled test trials yielded a significant intercept (ß=1.33, Z=4.2, $p<.001$), indicating above chance performance in the reference condition (Paradigm A, Block 1) and no main effect of block or paradigm, with only a marginal interaction (ß=.93, Z=1.8, $p=.073$); removing block and paradigm did not degrade model fit compared to the intercept-only model ($\chi^2$ (12)=5.74, $p=.93$). However, for novel-grammatical vs. novel-ungrammatical trials, the main effect of paradigm, block, and their interaction persisted. Again, these effects are driven by subjects' chance performance in Block 1, Paradigm A, (intercept ß=−.27, Z=−1.063, $p=.29$), whereas all other conditions are above chance. Specifically, subjects' performance in the novel-grammatical vs. ungrammatical trials in Block 1, Paradigm B was no different from their accuracy in repeated vs. scrambled trials in the same conditions (ß=.28, Z=.59, $p=.56$), and it was better than chance (intercept ß=.64, Z=2.36, $p=.018$). There were no main effect or interactions involving paradigm in Block 2. In Block 2, subjects selected novel-grammatical over ungrammatical strings at rates better than chance (intercept ß=.4773, Z=2.766, $p<.01$) corresponding to an accuracy of 61.5% . Subjects' accuracy in repeated vs. scrambled sentences in Block 2 was significantly better compared to novel-grammatical vs. novel-ungrammatical (ß=.541, Z=2.53, $p=.024$), corresponding to an accuracy of 73.5%.

That subjects performed better in repeated vs. scrambled trials is not surprising, as the that comparisons involves identical versions of training sentences versus strings in which both adjacent and non-adjacent sequences were disrupted—arguably a very easy comparison to make. In contrast, the novel-grammatical vs. novel-ungrammatical comparison involves two novel sentences whose surface properties are identical with respect to the training set, differing only in their conformity to more abstract properties of the training set, involving categories. However, even on those trials, we see above chance performance except for one subset of data in the first block.

### 2.3. Discussion

Subjects never experienced the novel-grammatical target words in the contexts in which they were tested—neither in the frames nor in the bigrams—yet they tended to choose them as words that belonged in those contexts, and endorsed novel-grammatical sentences. If they had based their judgments simply on the identity of the frame, then they should have accepted the novel-ungrammatical sentences as well, but they did not. The simplest explanation of this pattern of results is that exposure to the training material caused learners to form a category of target words within a paradigm, as a consequence of the target words' overlapping distributions. Learners then extended the co-occurrence privileges of target words to contexts in which they had not previously heard those words, but in which they had

heard other words from the same category, and consequently judged them to be more familiar. The block interaction, though limited to one paradigm, suggests that learners required extra exposure to learn the categories, compared to learning the more surface-level word sequence information.[2]

As mentioned previously, several recent studies have also shown that adults can form categories from distributional information, in the absence of correlated cues from other domains (Mintz, 2002; Reeder et al., 2013). This experiment thus provides further evidence that learners use distributional information alone to categorize words. In addition, we noted earlier that, as in this experiment, the distributional contexts in Mintz (2002) and Reeder et al. (2013) can be characterized as frames that surround the context words. Taken together, these findings raise the possibility that occurrence within a recurring frame was important for facilitating categorization and category-based generalizations of occurrence privileges. But before discussing this possibility further, we highlight one important difference between this experiment and the studies just cited: Here, the training material involved two distributional paradigms (i.e., two target categories), whereas in Mintz (2002) and Reeder et al. (2013), there was only one.

Differentiating novel-grammatical from novel-ungrammatical strings in the current study thus requires computations on two distinct sets of distributional information that are intermingled during the training phase. Hence, the current results provide new evidence regarding human learners' ability to generalize from distributional information. They show that even in a computationally more resource-demanding situation, when information about two distinct categories is present, learners can make category generalization using only distributional information.

We just speculated that learners' success in the current experiment and in Mintz (2002) and Reeder et al. (2013) may have been due to the presence of frames in the distributional patterns. However, as discussed earlier, scenarios in which frames could provide distributional information are scenarios in which the bigrams inherently contained in frames also could provide distributional information. It is possible that when learners categorized target words that co-occur within frames, they were attending only to the bigram patterns (between the target word and the immediately preceding word and immediately following word), and the fact that the framing words themselves frequently co-occur was irrelevant. In that case, learners should perform similarly in situations where the information provided from distributional patterns involving frames is severely diminished, but information provided by bigram patterns is intact. We test this hypothesis in Experiments 2 & 3. Each

[2]Prompted by a suggestion from an anonymous reviewer, we considered whether characteristics of the experimental design could have contributed to the improvement over blocks (for the subset of data in question). One possibility is that the juxtaposition of grammatical and ungrammatical items in the test trials after the first training block—e.g., grammatical: *swech lowfa klide*, ungrammatical: *swech puziv klide*—could have cued learners to notice the relevant distributional properties that distinguish them. Since such juxtapositions do not normally occur in children's input, if they were the cause of categorization it would limit what one could conclude from the categorization results. We find this possibility unlikely for two reasons. First, such a juxtaposition happens in only 4 of the 32 test trials. Second, presentation of a grammatical and ungrammatical item with the same frame gives no indication of which item is grammatical, only that grammaticality depends on the medial word. At best, then, the first set of test trials could provide a subtle cue to the location within sentences to which one should attend, but only the distributional information itself is informative about grammaticality. Combined with the fact that we only observed a block effect for the grammatical items in one paradigm, we do not think the first test session had an important impact on learners' categorization performance.

experiment was designed to hold constant different aspects of the training design structure in Experiment 1, while removing distributional information provided by frames. In Experiment 2, we simply withheld more items from the training set, such that any particular frame occurred with only one target word; the factors held constant was the fact that each target word was always preceded and followed by another word in each sentence, and bigram information was mostly predominantly matched with that in Experiment 1. In Experiment 3, we were able to perfectly match bigram information with Experiment 1, but we did so by changing global constraints on sentences, such that a target word could now begin or end a sentence.

## 3. Experiment 2

Experiment 2 assessed categorization using a subset of the training stimuli used in Experiment 1. In particular, the bigram patterns were similar to those in Experiment 1, but frames provided virtually no distributional information about categories. If learners used bigram information only when they formed distributional categories in Experiment 1, then they should show a similar categorization effect in Experiment 2, since the bigram information is similar across the two experiments. On the other hand, if the frame contexts were critical for categorization in Experiment 1, learners should not form categories in Experiment 2 because the distributional information provided by frames was considerably impoverished compared to Experiment 1.

### 3.1. Methods

**3.1.1. Subjects—**Fifteen college students participated in the experiment for partial satisfaction of course requirements. Subjects were randomly assigned to one of two counterbalanced training groups (see section 3.1.2). Two subjects in the first counterbalancing condition and one subject in the other were not included in the final analysis because they failed to reach a criterion of 50% correct on trails comparing repeated to scrambled sentences. Final data analysis thus included six subjects in each counterbalancing condition.

**3.1.2. Materials and Design—**The training set implemented bigram patterns that were similar to those in Experiment 1, but here, frames no longer provided robust distributional information about categories. We achieved this by selecting a subset of training sentences from Experiment 1 such that 12 of the 18 frames (i6 per paradigm) occurred with only one intervening target word. For example, the *choon X klide* frame only occurred in the sequence …*choon refton klide*…. However, the bigram patterns *choon X* (and *X klide*), nevertheless occurred with different target words in the *X* position. Specifically, there was a subset of four target words in the *choon X* bigram, and a different subset in the *X klide* bigram. In order to maintain the variability of target words in particular bigram patterns and within the same trigrams as in Experiment 1, it was necessary for the other six of the 18 frames (3 per paradigm) to occur with two target words, rather than just one. For example, the *choon X* bigram pattern was realized in the following frames: *choon refton klide*, *choon channer glaik*, *choon lartsue blit*, *choon lowfa glaik*; thus, *choon_glaik* occurred with two different target words.

To summarize, in Experiment 1, each frame occurred with four different target words, whereas in Experiment 2, most frames occurred with only one target word, but some occurred with two. Overall then, the distributional information provided by frames, while not completely absent, was considerably impoverished with respect to Experiment 1; On the other hand, bigram information was relatively stable across experiments: In Experiment 1, most target words (8) occurred with four context words, and some (4) occurred with six (including preceding and following contexts); in Experiment 2, every target word occurred in four contexts (see Table 1 and Table A.1).

Due to the differences in experimental design, there were 1/3 as many unique training sentences in Experiment 2 compared to Experiment 1, so each training sentence was repeated to make a total of three occurrences in the final training set. Thus, the number of training sentences was matched across experiments. In addition, the frequency of context words in bigrams was matched across experiments, each context word occurring 36 times across the three training blocks (i.e., occurring with 4 target words x 3 occurrences each with different combinations of optional words x 3 repetitions each = 36). Tables A.2 and A.3 summarize the bigram frequency counts for Experiments 1 & 2, respectively.

**3.1.3. Procedure—**The procedure was the same as in Experiment 1.

## 3.2. Results

We coded subjects' responses to the 2AFC questions as in Experiment 1. Average proportion correct across subjects for novel-grammatical versus novel-ungrammatical trials was 47.7% and for repeated versus scrambled trials was 72.7%. Figure 4 shows individual subject means for each comparison type. As in Experiment 1, we tested for a main effect and interaction of counterbalancing group with all within-subject variables (trial type, block, and paradigm) and their interactions. We used a mixed effect logistic regression, with the binary response score (correct=1, incorrect=0) as the dependent measure, controlling for random intercepts and slopes for the within-subject variables. There was no effect of counterbalancing, nor any interaction between counterbalancing condition and any other variables or their interactions. Therefore, we ran the same model without the counterbalancing variable, to increase power. There was a main effect of trial type (novel-grammatical/novel-ungrammatical vs. repeated/scrambled; ß=1.68, Z=4.97, $p<.001$), and no other main effects or interactions. A comparison of goodness of fit showed that the variables block and paradigm did not contribute significantly to model fit ($\chi^2$ (39)=24.2, $p=.97$). Table 3 shows the results from the model with trial type as the only variable, controlling for random subject variation on intercept and slope. The intercept of -.095 expresses the log-odds of correct over incorrect responses to the novel-grammatical vs. novel-ungrammatical 2AFC trails (corresponding to the 47.7% accuracy); this value was not significantly different from chance ($p=.42$), indicating that subjects did not generalize from the distributional patterns in the training set. However, the effect of trial type was significant ($p<.001$); subjects responded more accurately to the repeated versus scrambled 2AFC items compared to the novel-grammatical vs. novel-ungrammatical items (see Table 3). A separate analysis of the repeated versus scrambled trials, with subjects as a random effect, revealed that

subjects' responses to those items were also above chance (intercept ß=1.01, Z=6.353, $p<.001$).

Thus, while subjects were accurate at distinguishing verbatim repetitions of sentences from those in which the word order was scrambled, there was no evidence that subjects formed category generalization based on distribution of bigrams.

### 3.3. Discussion

The primary difference between Experiment 1 and Experiment 2 was in the nature of the distributional contexts from which learners could form category generalizations. In Experiment 1, both frames and bigrams entered into co-occurrence patterns with target words, whereas in Experiment 2, frames provided very little category information. In Experiment 2, subjects did not find novel-grammatical sentences to be more familiar than novel-ungrammatical sentences, indicating that they did not categorize the target words. The fact that learners categorized in Experiment 1, but not in Experiment 2, is consistent with the hypothesis that some property of the frames themselves facilitated generalization. When the available distributional information (almost) exclusively involved patterns of target words in bigrams, learners did not generalize lexical co-occurrence privileges.

It is interesting to note that Experiment 2 had similar design characteristics to an experiment in Reeder et al. (2013). In their Experiment 5b, target words occurred in frames, but every frame (except one) occurred exclusively with only one target word, similar to Experiment 2. As in Experiment 2, distributional information was carried primarily by bigram patterns. However, Reeder et al. found evidence of categorization whereas in Experiment 2 we did not. Despite the similarities between Reeder et al.'s Experiment 5b and our Experiment 2, there are important differences as well. As mentioned in section 1.2, subjects in the Reeder et al. study were exposed to distributional information pertaining to just one category, whereas here subjects were exposed to information about two. In other words, successful categorization in Reeder et al. amounted to distinguishing contexts from targets within one paradigm, whereas here, learners needed to acquire two distinct target categories from two distinct yet intermingled distributional paradigms. Moreover, in the experiments here, subjects were tested on a total of 12 target words (six per paradigm), whereas subjects in Reeder et al. (2013) were tested on three or four, depending on the experiment. Thus, our experiment arguably put greater demands on cognitive resources, but also may have engaged different mechanisms due to the different formal/computational properties of the stimuli. Frames may facilitate categorization when the distributional patterns are more complex. We discuss this topic further in the General Discussion.

Although we designed Experiment 2 to match the bigram information in Experiment 1, as mentioned in section 3.1.2, there were some differences across experiments in the details of the bigram patterns. In particular, all target words in Experiment 2 occurred in four bigram contexts—two involving the immediately preceding word, and two involving the immediately following word—whereas in Experiment 1, some target words occurred in six bigram contexts— three involving the immediately preceding word, and three involving the immediately following word. For example, compare Table A.2 and Table A.3, which summarize the bigram patterns in Experiments 1 & 2, respectively, and observe that the

target word *channer* is preceded by *ghire*, *choon*, and *swech* in Experiment 1, but only *choon* and *swech* in Experiment 2. In other words, although Experiment 2 did maintain substantial distributional information in bigram patterns, there was less somewhat less density and overlap in the set of bigram co-occurrence patterns for target words, compared to Experiment 1. Since that is just the kind of situation that Reeder et al. (Reeder et al., 2013) showed reduces generalization, it is conceivable that subjects' failure to categorize and generalize in Experiment 2 was due to this slightly reduced bigram information, rather than the (near) absence of distributional information from frames. If so, the results of Experiment 2 could not be taken as evidence for an advantage for frames over bigrams in facilitating category generalizations.

The slight differences in the bigram patterns in Experiment 2 arose in part from a constraint we imposed: As in Experiment 1, each occurrence of a target word involved a context word immediately to its right and left. This property, in combination with the overall constraint that the resulting frame should not itself provide relevant distributional information (with the exception of the three frames that occurred with two rather than one target word), resulted in the slightly different distribution of target words in bigrams.

In Experiment 3 we constructed a training set in which the bigram properties of Experiment 1 were duplicated exactly and in which frames provided no category information. To do this, we removed the constraint requiring that each token of a target word occur simultaneously with a preceding and following context. If learners treat the contexts preceding and following the target word independently—in other words, if they attend to bigrams but not frames—then it should not matter whether those contexts occur in the same sentence or in two different sentences (i.e., the preceding context in one, the following context in another). Learners attending only to bigrams should, therefore, categorize words equivalently in Experiments 1 & 3.

## 4. Experiment 3

Although the category information carried by bigram patterns in Experiment 2 was nearly identical to the bigram information in Experiment 1, there were some differences that, although small, could have diminished learners' ability to generalize from bigram information in Experiment 2. To address this, we designed the training sentences in Experiment 3 such that the patterns of target words in bigrams was identical to the patterns in Experiment 1, while removing completely the distributional patterns involving frames. In order to achieve this design constraint but also maintain the number of words per category, and other general properties of the language, target words in training sentences did not always occur simultaneously in two informative bigrams (defined by both the preceding and following word simultaneously), as was the case in Experiments 1 & 2. For example, rather than include …*choon sykteb klide*…, the training set included *choon sykteb*, and *sykteb klide* in different sentences (with filler words optionally appearing before *choon* and after *klide*). Since categorization from bigram information should not depend on having an informative preceding and following context *simultaneously* (which indeed is essentially a frame), the performance of a learning mechanism that is sensitive to the patterning of words within bigrams should not be degraded when the target words' participation in patterns with prior

and following words are decoupled. This experiment thus provides an additional means of testing whether learners in Experiment 1 could have generalized from bigram information without using information about the distributional patterns of words within frames.

## 4.1. Methods

**4.1.1. Subjects—**Fourteen college students participated in the experiment and received credit towards course assignments. Subjects were randomly assigned to one of two counterbalanced training groups (see section 4.1.2). Data for two subjects in one counterbalancing group were not included because the subjects failed to reach performance criterion on repeated/scrambled comparisons. Thus, data from 12 subjects (six in each counterbalancing training group) were included in the data analysis.

**4.1.2. Materials and Design—**As in the previous experiments, training sentences belonged to one of two paradigms, A and B, that defined the target word categories. Context and target words were the same as in the previous experiments. However, in this experiment, target words sometimes occurred flanked on each side by a context word as they did in the previous experiments, but sometimes they were flanked only on the left or only on the right. This allowed us to exactly match the bigram patterns from Experiment 1 that involved target words—matching both the items and the frequencies—while at the same time ensuring that frames provided no reliable category information for target words. As in the previous experiments, there were positions for optional words that occurred equally frequently in both paradigms, and thus from which no reliable category information could be computed. However, optional words never occurred immediately adjacent to a target word. For example, *channer blit* and *channer blit tiv* were both training sentences, but *dap channer blit* did not occur because *dap* is an optional word. This constraint ensured that target words entered into bigram patterns with exactly the same lexical items here as they did in Experiment 1.

As in the previous experiments, we devised a counterbalanced set of training sentences by switching the paradigm A and B target words in the novel-grammatical/ungrammatical test sentences—*lofa* and *pooda* were switched with *puziv*, and *mirshow*, respectively. The test trials were identical to those in Experiments 1 & 2. As a consequence, sentences that were novel-grammatical for one counterbalance group were novel-ungrammatical for the other.

There were 59 basic sentences—i.e., sentences without optional words—per paradigm, for a total of 118 basic sentences. (Because some training sentences only contained one bigram pattern as opposed to two, matching bigram frequency with Experiment 1 resulted in more training sentences than the in the prior experiments). The training set repeated each of the 118 sentences three times, with different combinations of optional words each time, resulting in 354 training sentences. Optional words never flanked a target word and thus were never in bigram patterns with target words. The bigram frequencies for this experiment and Experiment 1 are shown in Table A.2 in the Appendix.

The same periodic quiz questions were used during the training phrase; one question was administered every 19 training sentences.

**4.1.3. Procedure—**The procedure was the same as in the previous experiments.

## 4.2. Results

We coded subjects' responses to the 2AFC questions as in the previous experiments. Average proportion correct across subjects for novel-grammatical versus novel-ungrammatical trials was 49.7% and for repeated versus scrambled trials was 65.8%. Figure 5 shows individual subject means for each comparison type.

As in the previous experiments, we tested for a main effect and interaction of counterbalancing group with all within-subject variables (trial type, block, and paradigm) using a mixed effects logistic regression, with the binary response score (correct=1, incorrect=0) as the dependent measure, and controlling for random intercepts and slopes for the within-subject variables. There was no effect of counterbalancing, nor any interaction between counterbalancing condition and any other variables or their interactions. Therefore, we ran the same model without the counterbalancing variable, to increase power. There was no effect of paradigm or block, nor any interactions involving them; furthermore, those variables did not improve model fit ($\chi^2$ (39)=20.3, $p$=.99), so we performed an analysis with trial type (novel-grammatical/novel-ungrammatical or repeated/scrambled) as the only fixed effect, controlling for by-subject variance on the intercept and slope. Table 4 shows the results of the regression model. The intercept of (-.011) expresses the log-odds of correct over incorrect responses to the novel-grammatical versus novel- ungrammatical 2AFC trails, corresponding to the 49.7% accuracy; this value was not significantly different from chance ($p$=.93), indicating that subjects did not generalize from the distributional patterns in the training set. However, the effect of trial type was significant ($p$<.001); subjects responded more accurately to the repeated versus scrambled 2AFC items compared to the novel-grammatical versus novel-ungrammatical items. A separate analysis of the repeated versus scrambled trials, with subjects as a random effect, revealed that subjects' responses to those items were also above chance (intercept=.658, Z=6.115, $p$<.001).

Thus, as in Experiment 2, there was no evidence that subjects formed category generalization based on distribution of bigrams.

## 4.3. Discussion

Subjects in this experiment heard target words in the same bigram patterns as did subjects in Experiment 1, and they heard the bigrams with the same frequency in the two experiments. Nevertheless, subjects were equally likely to report novel-grammatical and novel-ungrammatical sentences as familiar in this experiment, whereas in Experiment 1 they were significantly more likely to endorse novel-grammatical sentences. The primary difference between the two experiments is that in Experiment 1, target words occurred within frequently occurring frames, whereas here, as in Experiment 2, they did not. Thus, it is unlikely that subjects' generalizations about target words in Experiment 1 were due exclusively to the bigram patterns involving the target words. Otherwise, subjects should have shown similar behavior in Experiment 3, where the bigram information was identical. Rather, the results of this experiment and Experiment 1 provide strong support for the hypothesis that learners based their generalizations on the distributional categories defined

by the frequently occurring frames in Experiment 1. While we cannot rule out the possibility that learners' categorization behavior responded to bigram patterns *in addition* to frames in Experiment 1, the results indicate that frames were a necessary component of the distributional information.

## 5. General Discussion

In this study we provide new evidence that human learners possess the mechanisms to categorize words using only distributional information, and without requiring top-down constraints from syntactic knowledge. This study goes further, however, by showing that human learners respond to certain distributional patterns more than others. In particular, learners categorized words when they occurred within frequently occurring frames (Experiment 1), but not when they occurred only within simpler bigram patterns (Experiments 2 & 3). These findings have exciting connections to recent computational studies that show that frequent frames at the word and morpheme level are informative distributional patterns cross-linguistically (Chemla et al., 2009; Mintz, 2003; Wang et al., 2011; Weisleder & Waxman, 2010). Thus, we have shown that the distributional contexts that provide accurate category information in natural languages are ones to which human learners appear biased to attend.

### 5.1. Frequent Frames From Linguistic and Cognitive Perspectives

Why would categorizing using frequent frames be beneficial for bootstrapping grammatical categories? Investigations of child-directed English by Wang & Mintz (2010 and in prep) suggest one possibility. They analyzed a parsed version of child-directed speech corpora that represented the grammatical structure of utterances using a relational grammar(Sagae, Davis, Lavie, Macwhinney, & Wintner, 2007) that linked each word in an utterance to another word and labeled the link with a grammatical relation (SUBJECT, OBJECT, etc.). For each frequent frame in a corpus, they analyzed the grammatical structures of each occurrence of the frame (as defined by the grammatical relations in which the words in the trigram entered). They found that, despite the potential variability across instances of the frequent frame, a large proportion of the instances (over 90% for most frames) occurred in very similar syntactic structures. This finding indicates that frequent frames occur in locations within an utterance that are syntactically highly homogeneous across occurrences, so that the target words that occur in the frame-medial position are considerably constrained with respect to the structure and therefore the grammatical category. In discussing this property, Wang & Mintz proposed that frequent frames might function as a proxy for structure-dependent distributional learning by allowing learners to identify regions in an utterance with linguistically informative distributional patterns. In other words, frequent frames essentially restrict distributional analyses to syntactically constrained domains, without requiring prior grammatical knowledge.

The finding that frequent frames select sequences with syntactic regularity naturally leads one to ask whether frequent frames select syntactic constituents, providing a basis for a hierarchical organization of words sequences. Indeed, frequent frames involving two pronouns are typically full transitive phrases, for example, *you_X_it* (Mintz, 2003). But

other frequent frames coincide with syntactic phrases only at one edge, for example *the_X_on*; while the *X* position is reliably a noun, the trigram selected by the *the_X_on* frame, and ones like it, do not coincide with traditional phrase structure configurations. Thus, the contribution of frequent frames might be most relevant with respect to grammatical categories (for a fuller discussion of the connection of frequent frames to linguistic structures, see Mintz, 2006b; Wang & Mintz, 2010).

In addition to the computational advantages offered by frequent frames over bigrams, trigrams characterized by frequent frames might offer some processing advantages as well. Carrying out distributional analyses necessarily involves evaluating particular elements (targets) with respect to a context. In processing the sequential information in speech, learners presumably do not know initially whether it would be more profitable, in terms of knowledge gain, to treat a given word as a context or a target (or neither). Information in the speech stream that could guide the learner in determining what elements to evaluate with respect to which other elements could greatly increase the informativeness and effectiveness of using distributional analyses to categorize words. From this perspective, it is not surprising that subjects used absolute word position as the categorization context in Smith (1966), where learners were exposed to two-word sequence. The sequences themselves did not provide any information to differentiate words as contexts or targets. Perhaps the frequently co-occurring framing elements in a frequent frame function to focus learners' attention on informative trigram sequences—the trigram sequences bounded by the frame elements—and the frame then acts as an anchor for analyzing the frame-medial words (see Valian & Coulson, 1988). At present, we have no independent evidence that bears on the mechanisms that orient learners towards frequent frames as distributional contexts. Additional studies are needed to develop a more complete account of why frequent frames are so readily detected and used by learners in category generalization, especially since detecting non-adjacent dependencies is argued to be difficult elsewhere the literature (e.g., Pacton & Perruchet, 2008 although their learning task was considerably different from ours). We discuss related issues with respect to infant learners in section 5.3.

### 5.2. The Role of Bigram Patterns In Lexical Categorization

It is important to emphasize that the present results do not show that human learners are incapable of using bigram patterns to categorize words. For instance, in Reeder et al. (2013), Experiments 5b & 5c, subjects were able to make category generalizations using bigram information. In those experiments, subjects were exposed to target words in frames, but, as in our Experiment 2, only bigrams could provide meaningful category information, since each frame occurred with only one target word. But, as we suggested in section 3.3, the resource demands were less in those experiments compared to the present study, and an advantage for frames might only emerge when materials in artificial languages are more complex. Moreover, in natural language—as opposed to some of the artificial languages just discussed (e.g., Smith, 1966)—function words that are likely to be the most informative in bigram patterns (e.g., determiners and auxiliary verbs) are very frequent; their high frequency could serve a similar filtering function of orienting learners towards distributionally informative locations, as we proposed for frequent frames—providing a salient anchor point in the utterance from which to analyze adjacent words (Braine, 1966;

Mintz et al., 2002; Valian & Coulson, 1988). The context words in Experiments 2 & 3 were also higher in frequency compared to the target words, and it is possible that with more exposure to the training materials, subjects would have formed generalizations based on bigram patterns. What the results show, however, is that learners generalized more readily from items within frequent frame environments compared to bigram environments alone.

In addition, these results do not provide any information about how sensitivities to different types of distributional information might change as the result of experience. For example, learners might initially use frequently occurring frames as contexts for forming category generalizations, but then notice other systematic distributional patterns after having categorized a number of words. These additional patterns could be ones that vary cross-linguistically, such as particularly informative context words in bigram patterns, sub-lexical morphemes, or converging cues from other domains (e.g., Monaghan, Chater, & Christiansen, 2005). Learners could then leverage these additional patterns and information sources for further learning. For example, studies with infants learning German (Hohle, Weissenborn, Kiefer, Schulz, & Schmitz, 2004) and French (Shi & Melançon, 2010) demonstrate that at 14 months of age, infants learning these languages can use the presence of a highly frequent function words in their respective language to categorize a following novel word. It is unknown, however, what the source of that distributional knowledge was. Since French and German both have informative frequent frames at the lexical or morphological level (Chemla et al., 2009; Stumper et al., 2011; Wang et al., 2011), one possibility is that infants used frames to carry out initial categorization and observed that the functors in question were highly diagnostic of the frame-based categories, and thus started to use them as additional distributional contexts (Wang et al., 2011). It is also possible that when frequency differences between words within bigrams are much greater than in our materials—the relative frequency of function words to content words in natural languages are generally much greater than in most artificial languages—learners start using very frequent words as generalization contexts for adjacent words, without first processing the frames in which they occur. In that case, the benefit for frequent frames that we see here may arise in situations in which individual context words do not surpass some relative frequency threshold. Further research is necessary to identify the factors that may influence the type of distributional information learners use. This, in turn, will advance our understanding of the processes by which infants develop language-specific categorization strategies.

Broadly, then, frequent frames may serve as a filter to focus learners on particular subsequences within an utterance—subsequences that contain particularly robust distributional information—but category generalizations might engage just a piece of the sequence (e.g., bigrams) once the distributional analyses mechanisms are focused on a particular section of the utterance. This is consistent with another finding in Wang & Mintz's (2010) analysis of the syntactic structures linked to frequent frames: They found that in the trigrams defined by frequent frames in English, the first and second word positions (i.e., the first bigram) very often involved words that were linked by a grammatical relationship (e.g., SUBJECT OF), and almost exclusively the same grammatical relation for all tokens of a particular frequent frame. The consequence is that the first bigram within a frequent frame offers highly accurate information about the category of the target word.

Importantly, Wang & Mintz also analyzed bigrams simply defined as frequent words (unigrams) as the context for immediately adjacent target words. In those bigrams that were not constrained to be internal to frequent frames, there were significantly fewer instances of grammatical relationships between the two words; moreover, when the words were related, there was considerably more variability in the types of relations. In other words, bigrams within frequent frames were much more syntactically constrained than bigrams overall.

Recall from section 1.1 that St. Clair et al. (St. Clair et al., 2010) proposed that a superior method of categorizing from trigrams is to consider the two simultaneously occurring bigrams that form the trigram, rather than the frame. While, their proposal is consistent with the findings just mentioned regarding the informativeness of bigrams within frequent frames (Wang & Mintz, 2010), our findings provide no evidence that learners were, in fact, using these simpler bigram patterns to categorize words. If they were, we would have expected to find evidence of categorization in Experiments 2 & 3, where bigram information was informative of category membership but frames were not; yet we found no such evidence. On the other hand, the flexibility in representations provided by the flexible frames approach is intuitively desirable, as it seems unlikely that learners would represent a context word *only* as part of a frame, as opposed to as an independent word. Yet our results show that there are situations in which, with respect to *categorization*, learners initially base their generalizations on the context provided by frames.

### 5.3. Generalizing To Infant Categorization Mechanisms

Although subjects in these experiments were adults, the broader goal of this study was to advance our understanding of the types of information that human infants use when they initially start to categorize words. Our operating assumption was that the generalization mechanisms we are investigating are fundamentally the same between infants and adults, although the parameters that control these mechanisms are likely to be different (Hudson Kam & Newport, 2005). Indeed, similar assumptions are shared, implicitly or explicitly, in many studies of artificial language learning with adults. Nevertheless, the differences in representations and processes available to infants and adults with respect to category generalization might be important. Indeed, part of the argument advanced by St. Clair et al. (2010) in favor of bigrams and against frames as an early source of category information was that infants may not be able to detect and store frames early on. To represent a frame, learners must detect the non-adjacent dependency between the two framing elements. Gómez & Maye (2005) reported that infants exposed to artificial language materials can detect deviations from learned non-adjacent dependencies at 15 months of age, but not at 12 months. St. Clair et al. argued that those findings, along with evidence that eight-month-olds can detect adjacent dependencies (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996), is most consistent with an early reliance on bigram information for categorization, not frames. While the behavioral results we report here provide counter evidence to this hypothesis with respect to adults, they do not directly address the question with respect to infants. Recall from the previous section that studies with 14-month-olds learning German (Hohle et al., 2004) and French (Shi & Melançon, 2010) found that infants can generalize from some bigram patterns; if the account we proposed is correct—i.e., that frequent frames are the information filter that help learners detect particularly informative

words to anchor bigram analyses—then infants must be able to detect these patterns before 14 months of age. It is noteworthy that a recent study established that infants as young as seven months can detect non-adjacent repetition patterns (e.g., *bi-la-bi*; Gervain & Werker, 2013), suggesting that infants may process and represent non-adjacent relationships early enough for them to be used in initial category learning. However, the repetition of elements has been argued to have a heightened perceptual salience (e.g., Endress, Nespor, & Mehler, 2009), so one must be cautious in generalizing those findings to frequent frames, which typically involve dependencies between different words. Moreover, the non-adjacent dependencies tested by Gervain & Werker were between syllables in continuous speech, not between words. Nevertheless, in an analysis of two large corpora of English infant-directed speech, we found that approximately 83% of word tokens were monosyllabic.[3] Hence, there is some justification in generalizing findings pertaining to infants' detection of patterns within sequences of syllables to capabilities they may have with respect to detecting patterns within word sequences, at least in infant-directed speech. Relatedly, it is worth noting that the experiment that failed to find evidence of non-adjacent dependency learning in 12-month-olds (Gomez & Maye, 2005) used stimuli in which (as here) there were bisyllabic medial words, and words were separated by brief pauses. It could be that younger infants are aided by continuous speech and monosyllabic words in detecting non-adjacent patterns.

In summary, while there are some hints in the literature that infants may be able to detect the kinds of distributional patterns that are essential to our proposal, there is presently no strong evidence that they can. A study like the present one, but with infants, would provide the most direct test of our proposal. With that aim, we have begun to examine this question with 12- and 15-month-old infants.

## 6. Conclusion: A Generalized Account Reconciling Present Findings With Prior Research

One way of viewing the potential role of frequent frames in early word categorization is that they could focus the learner on sections of an utterance that are particularly informative with respect to distributional information. The frame itself can then be used to categorize the target words, or the learner could use a bigram pattern within the frequent frame. In either case, the frame can be viewed as playing a similar role as syntactic knowledge in structure-dependent distributional learning (Pinker, 1984)—constraining the distributional analysis to particularly informative regions of an utterance.

This view of frequent frames offers a perspective on the prior studies that concluded that learners require converging cues from other domains in order to make use of distributional information. As we mentioned in section 1, those studies involved bigram patterns in artificial languages. According to our hypotheses, bigrams alone may have been inadequate for stimulating distributional learning, in part because of the inherent difficultly of identifying the relevant patterns in the first place, and identifying the contexts and targets

[3]We analyzed two corpora of infant-directed speech in the CHILDES database (Macwhinney, 2000) using a syllabified version of the CMU Pronouncing Dictionary (http://webdocs.cs.ualberta.ca/~kondrak/cmudict/cmudict.rep; Bartlett, Kondrak, & Cherry, 2009) to derive syllable counts for all word tokens directed to children. Approximately 85% of child-directed tokens in the Bernstein-Ratner corpus (Bernstein-Ratner, 1987) and 83% in the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) were monosyllabic.

within those patterns. Correlated cues could facilitate this process. Just as we argued that the co-occurring framing elements of frequent frames could serve to focus attention on relevant sequences, the marking of certain words with consistent phonological or semantic information (Braine, 1987; Frigo & McDonald, 1998; Gomez & Lakusta, 2004) could serve a similar role of focusing attention.

On this view, learners can make use of distributional information to categorize words provided that they are given guidance as to what sequences to analyze, and some hints about how to analyze them (i.e., what to treat as contexts and targets). Structure-dependent distributional learning (Pinker, 1984) was conceived on theoretical grounds to fill this role, but learners may well be poised to make use of more indirect, yet computationally simple links to structure as a way of constraining their distributional analyses. Correlated cues from other domains may play a role in language-specific ways, once distributional learning has provided some structural organization. Here we found situations in which learners can use distributional information alone to make generalizations about word categories, but only when the distributional information was signaled by frames, not when it was signaled only by bigrams.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-0721328. Research reported in this publication was also supported by the National Institute of Child Health and Human Development of the National Institutes of Health under award number R01HD040368. We would like to thank Laura Siegel and Elsi Kaiser for their input on earlier drafts of this manuscript.

## References

Arnon I, Ramscar M. Granularity and the acquisition of grammatical gender: how order-of-acquisition affects what gets learned. Cognition. 2012; 122(3):292–305. [PubMed: 22169657]

Aslin RN, Saffran JR, Newport EL. Computation of Conditional Probability Statistics by 8- Month-Old Infants. Psychological Science. 1998; 9(4):321–324.

Bartlett, S.; Kondrak, G.; Cherry, C. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics; 2009 May. On the syllabification of phonemes; p. 308-316.

Bernstein-Ratner, N. The Phonology of Parent-Child Speech. In: Nelson, K.; van Kleeck, A., editors. Children's language. Vol. 6. Hillsdale, NJ: Lawrence Erlbaum Associates; 1987.

Bloomfield, L. Language. New York: Holt, Reinehart, and Winston; 1933.

Boersma, P.; Weenink, D. Praat: doing phonetics by computer (Version 5.1.43)[Computer program]. 2009. Retrieved from http://www.praat.org/

Braine MD. Learning the positions of words relative to a marker element. Journal of Experimental Psychology. 1966; 72(4):532–540. [PubMed: 5969727]

Braine, MDS. What is learned in acquiring word classes: A step toward an acquisition theory. In: Macwhinney, B., editor. Mechanisms of Language Acquisition. Hillsdale, NJ: Lawrence Erlbaum Associates; 1987. p. 65-88.

Braine MDS, Brody RE, Brooks PJ, Sudhalter V, Ross JA, Catalano L, Fisch SM. Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. Journal of Memory and Language. 1990; 29(5):591–610.

Cartwright TA, Brent MR. Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. Cognition. 1997; 63(2):121–170. [PubMed: 9233082]

Chemla E, Mintz TH, Bernal S, Christophe A. Categorizing words using “frequent frames”: what cross-linguistic analyses reveal about distributional acquisition strategies. Developmental Science. 2009; 12(3):396–406. [PubMed: 19371362]

Cohen JD, Macwhinney B, Flatt M, Provost J. PsyScope: A new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments, and Computers. 1993; 2:257–271.

Endress AD, Nespor M, Mehler J. Perceptual and memory constraints on language acquisition. Trends in Cognitive Sciences. 2009; 13(8):348–353. [PubMed: 19647474]

Erkelens, M. Learning to categorize verbs and nouns: Studies on Dutch. Utrecht: LOT; 2009.

Fisher C, Hall DG, Rakowitz S, Gleitman LR. When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. Lingua. 1994; 92:333–375.

Frigo L, McDonald J. Properties of phonological markers that affect the acquisition of gender-like subclasses. Journal of Memory and Language. 1998; 39(2):218–245.

Gerken L, Wilson R, Lewis W. Infants can use distributional cues to form syntactic categories. Journal of Child Language. 2005; 32(02):249–268. [PubMed: 16045250]

Gervain J, Werker JF. Learning non-adjacent regularities at age 0 ; 7. Journal of Child Language. 2013; 40(4):860–872. [PubMed: 22863363]

Gleitman LR. The Structural Sources of Verb Meanings. Language Acquisition. 1990; 1(1):3–55.

Gleitman LR, Gleitman H. A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. Current Directions in Psychological Science. 1992; 1(1):31–35.

Gleitman LR, Cassidy K, Nappa R, Papafragou A, Trueswell JC. Hard Words. Language Learning and Development. 2005; 1(1):23–64.

Gomez RL. Variability and detection of invariant structure. Psychological Science. 2002; 13(5):431–436. [PubMed: 12219809]

Gomez RL, Gerken L. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. Cognition. 1999; 70(2):109–135. [PubMed: 10349760]

Gomez RL, Lakusta L. A first step in form-based category abstraction by 12-month-old infants. Developmental Science. 2004; 7(5):567–580. [PubMed: 15603290]

Gomez RL, Maye J. The Developmental Trajectory of Nonadjacent Dependency Learning. Infancy. 2005; 7(2):183–206.

Harris, ZS. Structural Linguistics. Chicago: University of Chicago Press; 1951.

Hohle B, Weissenborn J, Kiefer D, Schulz A, Schmitz M. Functional elements in infants’ speech processing: The role of determiners in the syntactic categorization of lexical elements. Infancy. 2004; 5(3):341–353.

Hudson Kam C, Newport EL. Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. Language Learning and Language Development. 2005; 1(2):151–195.

Jaeger TF. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. Journal of Memory and Language. 2008; 59(4):434–446. [PubMed: 19884961]

Jordan, J. Shntool. [Computer Program]. 2009 Mar 30. Retrieved from http://shnutils.freeshell.org

Macwhinney, B. The Database. 3rd Edition. Vol. 2. Mahwah, NJ: Lawrence Erlbaum Associates; 2000. The CHILDES Project: Tools for analyzing talk.

Maratsos, MP. The child's construction of grammatical categories. In: Wanner, E.; Gleitman, LR., editors. Language Acquisition: The State of the Art. Cambridge: Cambridge Univ Press; 1982. p. 240-266.

Maratsos, MP.; 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.

Mintz TH. Category induction from distributional cues in an artificial language. Memory & Cognition. 2002; 30(5):678–686. [PubMed: 12219885]

Mintz TH. Frequent frames as a cue for grammatical categories in child directed speech. Cognition. 2003

Mintz, TH. Finding the verbs: distributional cues to categories available to young learners. In: Golinkoff, RM.; Hirsh-Pasek, K., editors. Action meets word: How children learn verbs. New York: Oxford University Press; 2006a.

Mintz, TH. Frequent Frames: Simple Co-occurrence Constructions and Their Links to Linguistic Structure. In: Clark, EV.; Kelly, BF., editors. Constructions in Acquisition. Stanford, CA: CSLI Publications; 2006b.

Mintz, TH. Fifteen-month-old infants can categorize words using distributional information alone and retain the categories after 1 week; Presented at the 36th Annual Boston University Conference on Language Development; Boston. 2011. Retrieved from http://www.frontiersin.org/people/TobenMintz/25123/video

Mintz, TH.; Newport, EL. Distributional regularities of form class in speech to young children; Presented at the PROCEEDINGS-NELS; 1995.

Mintz TH, Newport EL, Bever TG. The distributional structure of grammatical categories in speech to young children. Cognitive Science. 2002; 26(4):393–424.

Monaghan P, Chater N, Christiansen MH. The differential role of phonological and distributional cues in grammatical categorisation. Cognition. 2005; 96(2):143–182. [PubMed: 15925574]

Monaghan P, Christiansen MH, Chater N. The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. Cognitive Psychology. 2007; 55(4):259–305. [PubMed: 17291481]

Pacton S, Perruchet P. An attention-based associative account of adjacent and nonadjacent dependency learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2008; 34(1):80–96.

Pinker, S. Language Learnability and Language Development. Cambridge, MA: Harvard University Press; 1984.

Redington M, Chater N, Finch S. Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science. 1998; 22(4):425–469.

Reeder PA, Newport EL, Aslin RN. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. Cognitive Psychology. 2013; 66(1):30–54. [PubMed: 23089290]

Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science. 1996; 274(5294):1926–1928. [PubMed: 8943209]

Sagae K, Davis E, Lavie A, Macwhinney B, Wintner S. High-accuracy annotation and parsing of CHILDES transcripts. Proceedings of ACL Workshop on Cognitive Aspects of Computational Language Acquisition. 2007:25–32.

Shi R, Melançon A. Syntactic Categorization in French-Learning Infants. Infancy. 2010; 15(5):517–533.

Shi R, Morgan JL, Allopenna P. Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. Journal of Child Language. 1998; 25(1):169–201. [PubMed: 9604573]

Smith KH. Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? Journal of Experimental Psychology. 1966; 72(4):580–588. [PubMed: 5969733]

Smith KH. Learning Co-occurrence restrictions: Rule induction or rote learning? Journal of Verbal Learning and Verbal Behavior. 1969; 8(2):319–321.

St. Clair MC, Monaghan P, Christiansen MH. Learning grammatical categories from distributional cues: flexible frames for language acquisition. Cognition. 2010; 116(3):341–360. [PubMed: 20674613]

Stumper B, Bannard C, Lieven E, Tomasello M. “Frequent frames” in German child-directed speech: a limited cue to grammatical categories. Cognitive Science. 2011; 35(6):1190–1205. [PubMed: 21790746]

Theakston AL, Lieven EV, Pine JM, Rowland CF. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. Journal of Child Language. 2001; 28(1):127–152. [PubMed: 11258001]

Tomasello M. The item-based nature of children's early syntactic development. Trends in Cognitive Sciences. 2000; 4(4):156–163. [PubMed: 10740280]

Valian V, Coulson S. Anchor points in language learning: The role of marker frequency. Journal of Memory and Language. 1988; 27(1):71–86.

Wang, H.; Mintz, TH. A Supllement to the Proceedings of the 34th annual Boston University Conference on Language Development. Somerville, MA: Cascadilla Press; 2010. From linear sequences to abstract structures: Distributional information in infant-directed speech.

Wang, H.; Hohle, B.; Ketrez, FN.; Kuntay, AC.; Mintz, TH. Cross-linguistic Distributional Analyses with Frequent Frames: The Cases of German and Turkish. In: Danis, N.; Mesh, K.; Sung, H., editors. Proceedings of the 35th annual Boston University Conference on Language Development. Somerville, MA: Cascadilla Press; 2011. p. 628-640.

Weisleder A, Waxman SR. What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English. Journal of Child Language. 2010; 37(5):1089–1108. [PubMed: 19698207]

## Appendix

**Table A.1.**

Trigrams used in the training set for paradigm B in Experiment 1 (E1) and Experiment 2 (E2). Rows indicate the first and last words (frame) and columns represent the middle words (target). An E1 or E2 in the cell indicates that the designated trigram occurred in the materials for the experiment indicated.

| | **Plosit** | **lifik** | **puziv** | **antow** | **grimpot** | **mirshow** |
|---|---|---|---|---|---|---|
| jub_fex | E1 | E1 | E1 | | | E1,E2 |
| jub_sowch | E1 | | E1,E2 | E1,E2 | E1 | |
| jub_voy | | E1 | | E1 | E1,E2 | E1 |
| Dess_fex | E1,E2 | E1 | E1,E2 | | E1 | |
| Dess_sowch | | E1,E2 | E1 | E1 | E1 | |
| dess_voy | E1 | E1 | | E1,E2 | E1 | |
| ghip_fex | E1 | E1,E2 | | E1 | | E1 |
| ghip_sowch | E1 | E1 | | E1 | E1,E2 | |
| ghip_voy | E1,E2 | | | E1 | E1 | E1,E2 |

**Table A.2.**

Frequency counts for target-final bigrams (Table a) and target-initial bigrams (Table b) in Experiments 1 & 3. Counts are for one of the three familiarization blocks in paradigm A. Columns pertain to a particular target word, rows pertain to a particular context word. The frequency counts for the other two blocks in the paradigm are identical, the only difference in blocks being the flanking words (not indicated, and never adjacent to a target). The structure is identical for paradigm B, but with different words.

**(a)**

| | **lowfa** | **sykteb** | **lartsue** | **channer** | **refton** | **pooda** | **Σ*targets*** |
|---|---|---|---|---|---|---|---|
| Swech | 0 | 3 | 3 | 2 | 2 | 4 | *14* |
| ghire | 4 | 2 | 2 | 2 | 2 | 4 | *16* |
| Choon | 4 | 2 | 2 | 3 | 3 | 0 | *14* |

**(a)**

| | **lowfa** | **sykteb** | **lartsue** | **channer** | **refton** | **pooda** | **Σ*targets*** |
|---|---|---|---|---|---|---|---|
| Σ*contexts* | *8* | *7* | *7* | *7* | *7* | *8* | **44** |

**(b)**

| **lowfa** | **sykteb** | **lartsue** | **channer** | **refton** | **pooda** | **Σ*targets*** | |
|---|---|---|---|---|---|---|---|
| 0 | 3 | 2 | 3 | 2 | 4 | *14* | klide |
| 4 | 2 | 2 | 2 | 2 | 4 | *16* | glaik |
| 4 | 2 | 3 | 2 | 3 | 0 | *14* | blit |
| *8* | *7* | *7* | *7* | *7* | *8* | **44** | Σ*contexts* |

**Table A.3.**

Frequency counts for target-final bigrams (Table a) and target-initial bigrams (Table b) in Experiment 2. Counts are for one of the three familiarization blocks in paradigm A. Columns pertain to a particular target word, rows pertain to a particular context word. The frequency counts for the other two blocks in the paradigm are identical, the only difference in blocks being the flanking words (not indicated, and never adjacent to a target). The structure is identical for paradigm B, but with different words.

**(a)**

| | **Lowfa** | **sykteb** | **lartsue** | **channer** | **refton** | **pooda** | **Σ*targets*** |
|---|---|---|---|---|---|---|---|
| Swech | 0 | 3 | 3 | 3 | 0 | 3 | *12* |
| ghire | 3 | 3 | 0 | 3 | 3 | 3 | *12* |
| Choon | 3 | 0 | 3 | 3 | 3 | 0 | *12* |
| Σ*contexts* | *6* | *6* | *6* | *6* | *6* | *6* | **36** |

**(b)**

| **lowfa** | **sykteb** | **lartsue** | **channer** | **refton** | **pooda** | **Σ*targets*** | |
|---|---|---|---|---|---|---|---|
| 0 | 3 | 0 | 3 | 3 | 3 | *12* | klide |
| 3 | 0 | 3 | 3 | 0 | 3 | *12* | glaik |
| 3 | 3 | 3 | 0 | 3 | 0 | *12* | blit |
| *6* | *6* | *6* | *6* | *6* | *6* | **36** | Σ*contexts* |

## Highlights

- We investigate how language learners categorize words.
- We compare learning from two kinds of distributional patterns: Frames and Bigrams.
- We found that learners categorized using frames, but not bigrams.
- The patterns to which learners attend are informative cross-linguistically.

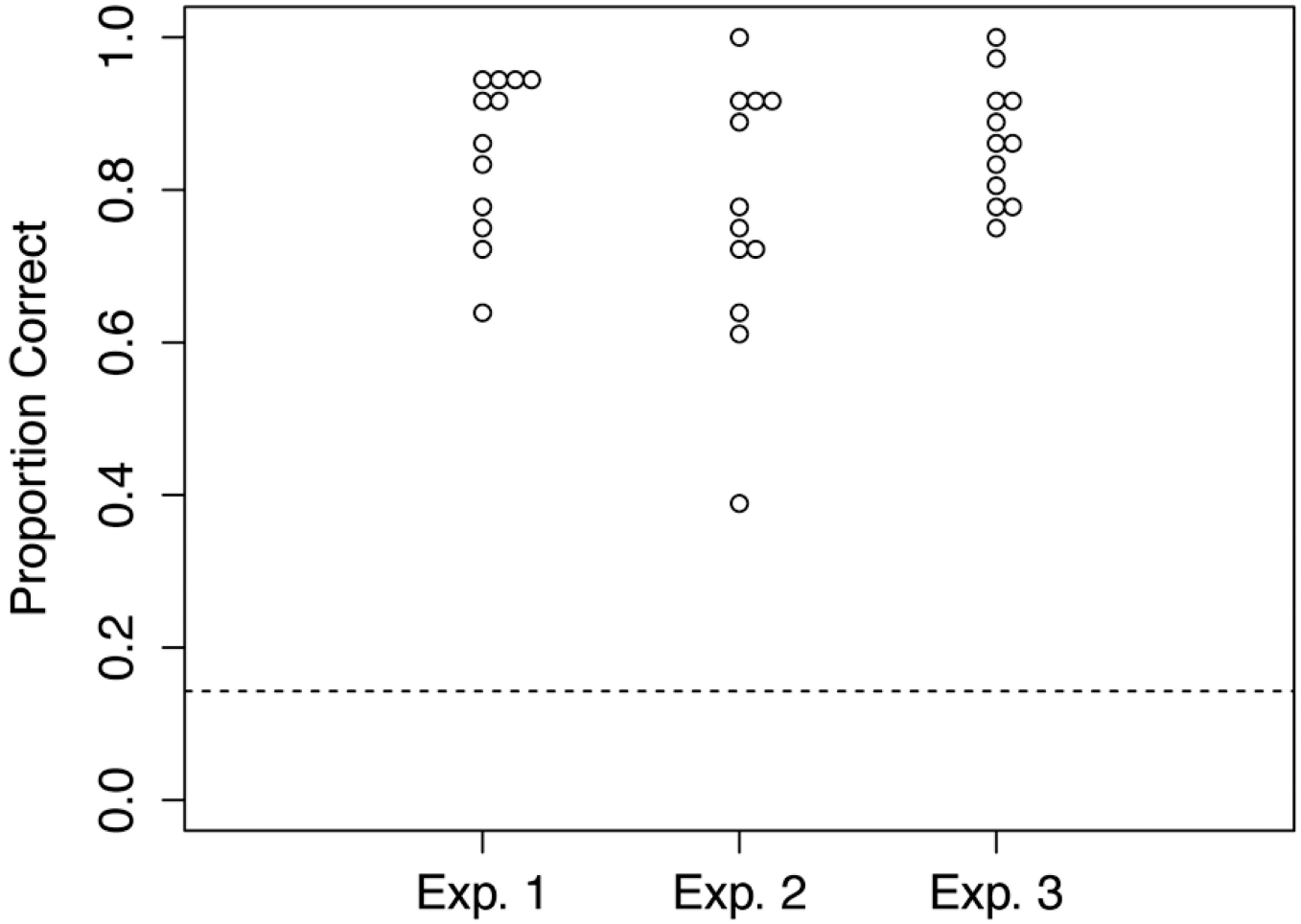


**Figure 1.**

Quiz accuracy per subject across Experiments 1–3. Dashed line indicates chance proportion correct. All subjects performed significantly above chance.

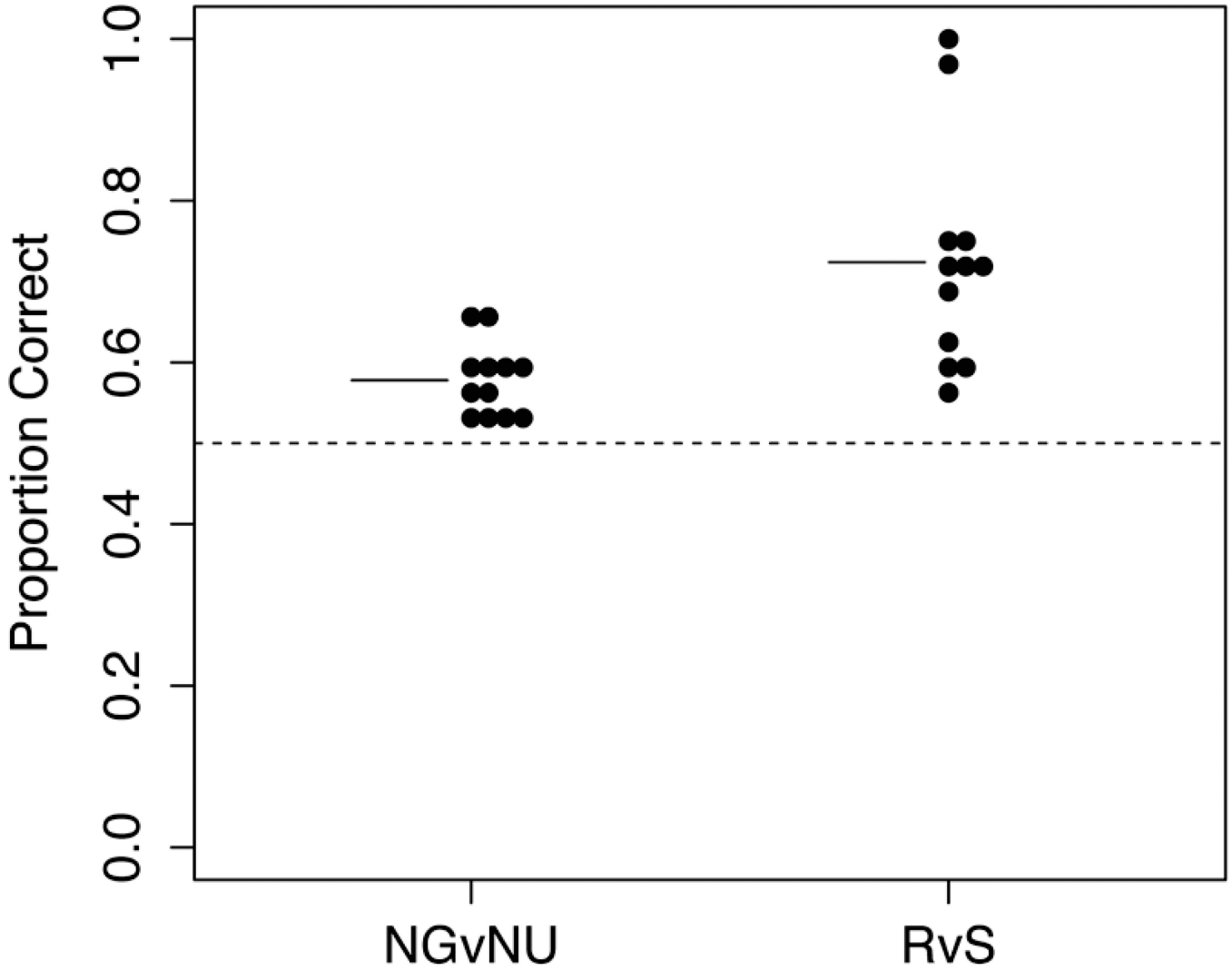


**Figure 2.**

Performance on 2AFC trials in Experiment 1, organized by trial type (NGvNU = novel-grammatical versus novel-ungrammatical, RvS = repeated versus scrambled). Mean proportion correct for each condition is indicated by a solid line. Chance (50%) is indicated by the dashed line.

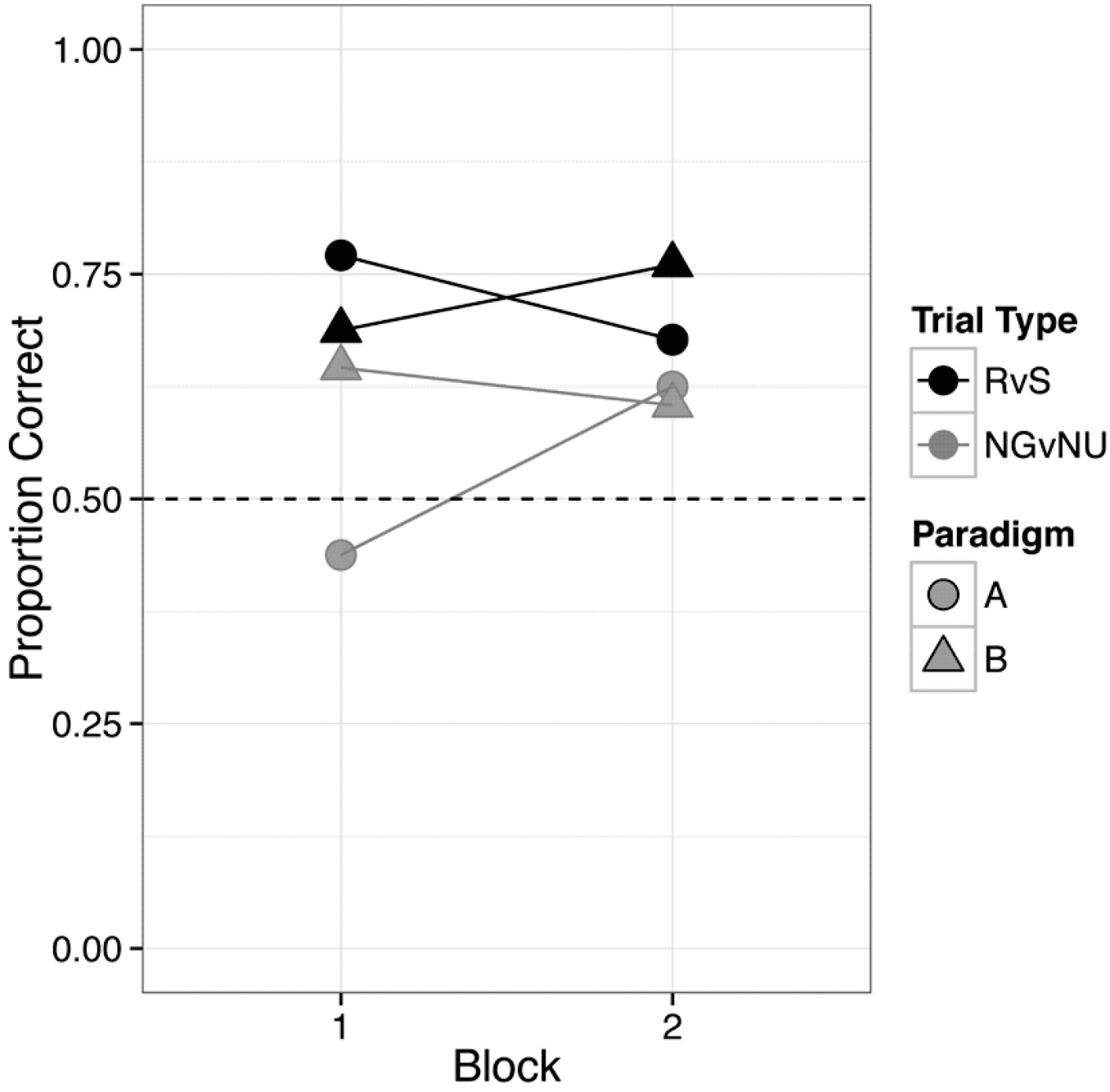


**Figure 3.**

Performance on 2AFC in Experiment 1. Mean proportion correct, averaged across subjects and broken down by condition (RvS = repeated vs. scrambled, NGvNU = novel-grammatical vs. novel-ungrammatical). The dashed line depicts chance performance (50% correct).

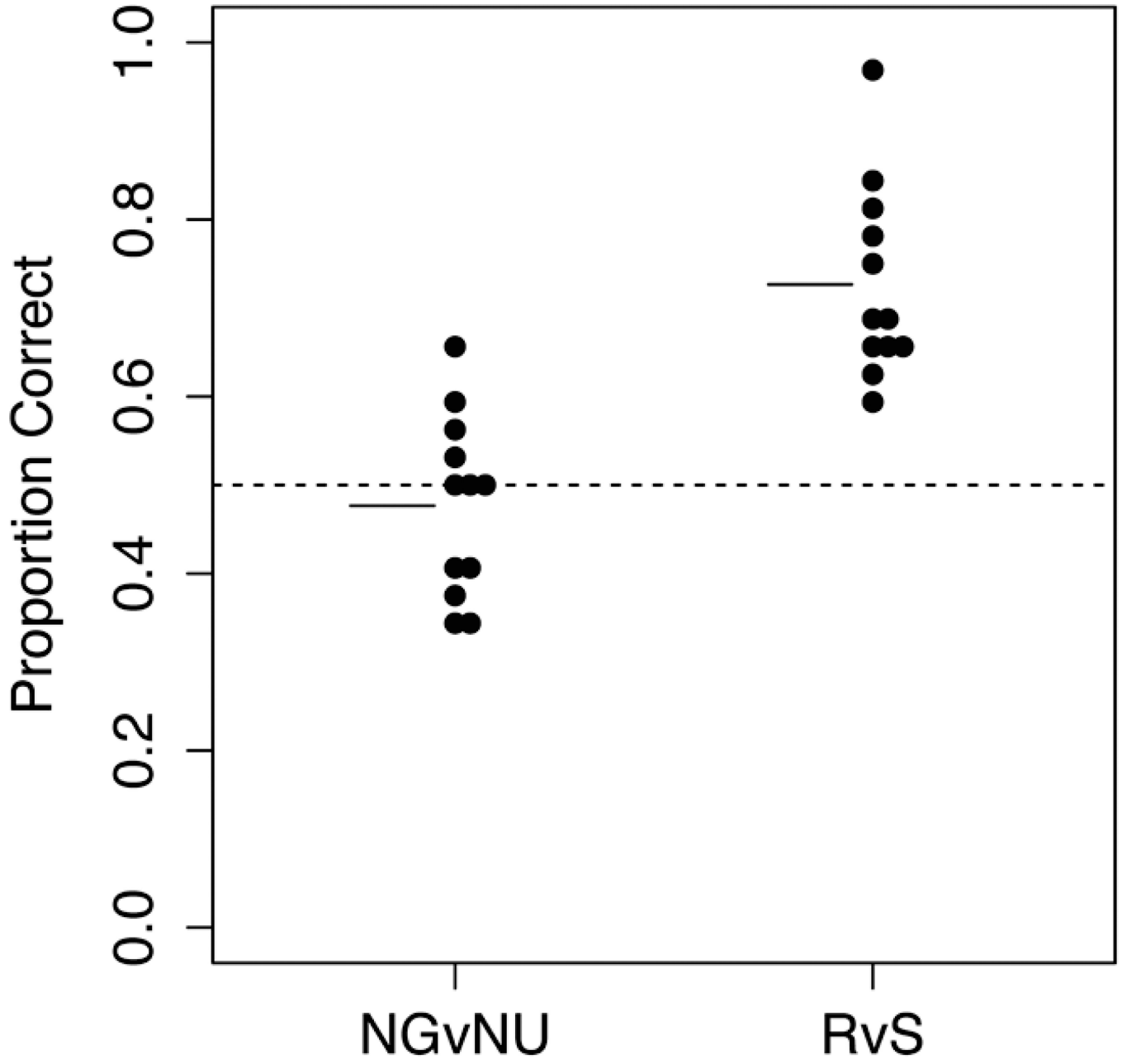


**Figure 4.**

Performance on 2AFC trials in Experiment 2, organized by trial type (novel-grammatical versus novel-ungrammatical = NGvNU, repeated versus scrambled = RvS). Mean %-correct for each condition is indicated by a solid line. Chance (50%) is indicated by the dashed line.

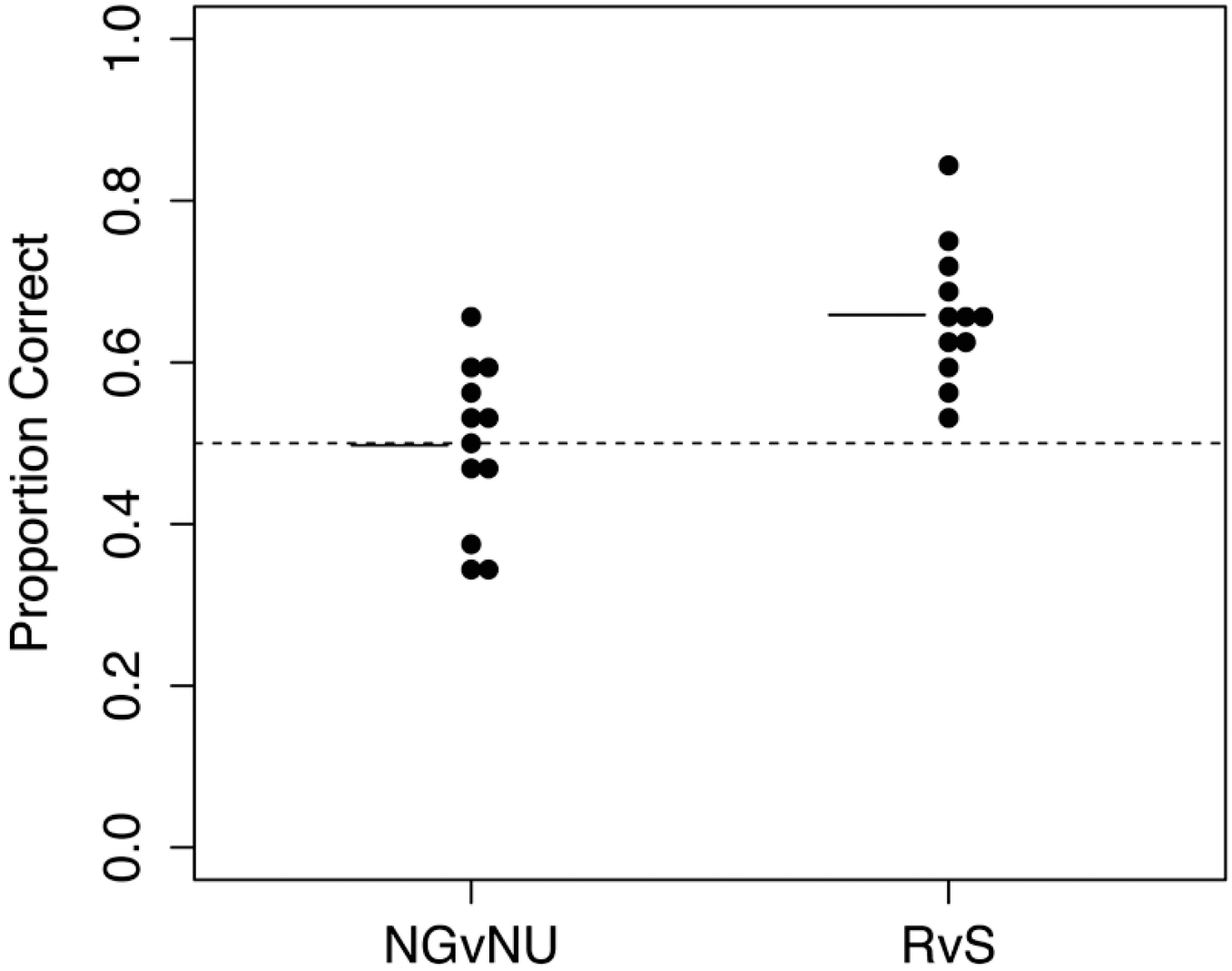


**Figure 5.**

Performance on 2AFC trials in Experiment 3, organized by trial type (NGvNU = novel-grammatical versus novel-ungrammatical, RvS = repeated versus scrambled). Mean %-correct for each condition is indicated by a solid line. Chance (50%) is indicated by the dashed line.

**Table 1**

Trigrams used in the training set for paradigm A in Experiment 1 (E1) and Experiment 2 (E2). Rows indicate the first and last words (frame) and columns represent the middle words (target). An E1 or E2 in the cell indicates that the designated trigram occurred in the materials for the experiment indicated. The structure of paradigm B was identical, but involved a completely different set of words (see Table A.1. in the Appendix).

| | **channer** | **lartsu** | **lowfa** | **refton** | **sykteb** | **pooda** |
|---|---|---|---|---|---|---|
| ghire_glaik | E1 | E1 | E1 | | | E1,E2 |
| ghire_blit | E1 | | E1,E2 | E1,E2 | E1 | |
| ghire_klide | | E1 | | E1 | E1,E2 | E1 |
| choon_glaik | E1,E2 | E1 | E1,E2 | | E1 | |
| choon_blit | | E1,E2 | E1 | E1 | E1 | |
| choon_klide | E1 | E1 | | E1,E2 | E1 | |
| swech_glaik | E1 | E1,E2 | | E1 | | E1 |
| swech_blit | E1 | E1 | | E1 | E1,E2 | |
| swech_klide | E1,E2 | | | E1 | E1 | E1,E2 |

## Table 2

Results of the logistic regression model in Experiment 1, including all within-subjects variables as fixed effects, with by-subject intercepts and slopes as random effects. The intercept reflects aggregate performance in the reference condition: novel-grammatical vs. novel-ungrammatical, Block 1, Paradigm A. NGvNU = novel-grammatical vs. novel-ungrammatical, RvS = repeated vs. scrambled.

| | **Estimate** | **Std. Error** | **Wald's Z** | ***p*-value** |
|---|---|---|---|---|
| Intercept (reference condition: NGvNU, Paradigm A, Block 1) | −0.2513 | .2057 | −1.222 | .22189 |
| Trial Type (NGvNU vs. RvS) | 1.6042 | .3906 | 4.107 | **$4.01\times10^{-5}$** |
| Block (1 vs. 2) | 0.7700 | .3031 | 2.541 | **.01106** |
| Paradigm (A vs. B) | 0.8947 | .3495 | 2.560 | **.01047** |
| Trial Type X Block | −1.2871 | .4542 | −2.834 | **.00460** |
| Trial Type X Paradigm | −1.2704 | .5158 | −2.463 | **.01378** |
| Block X Paradigm | −0.9709 | .4806 | −2.020 | **.04335** |
| Trial Type X Block X Paradigm | 1.8828 | .6681 | 2.818 | **.00483** |

## Table 3

Results of the logistic regression model. The intercept indicates subjects were at chance in choosing novel-grammatical sentences over novel-ungrammatical sentences. The significant trial type effect demonstrates that subjects were more likely to choose repeated over scrambled sentences than they were to choose novel-grammatical over novel-ungrammatical.

| | Estimate | Std. Error | Wald's z | *p*-value |
|---|---|---|---|---|
| Intercept (novel-grammatical vs. novel-ungrammatical) | −.0946 | .1177 | −.804 | .421 |
| Trial Type (contrasting repeated vs. novel against intercept) | 1.1045 | .1632 | 6.767 | **$1.3\times10^{-11}$** |

## Table 4

Results of logistic regression in Experiment 3. The intercept indicates subjects were at chance in choosing novel-grammatical sentences over novel-ungrammatical sentences. The significant trial type effect demonstrates that subjects were more likely to choose repeated over scrambled sentences than they were to choose novel-grammatical over novel-ungrammatical.

| | **Estimate** | **Std. Error** | **Wald's z** | ***p*-value** |
|---|---|---|---|---|
| Intercept (novel-grammatical vs. novel-ungrammatical) | −.01053 | .11512 | −.091 | .927 |
| Trial Type (contrasting repeated vs. novel against intercept) | .67053 | .15004 | 4.469 | **$7.9\times10^{-6}$** |