Language learners produce novel relative clause types without exposure: Evidence for a single, general syntactic representation of long-distance dependencies

Adam M. Morgan[1] & Victor S. Ferreira[2]

[1] NYU School of Medicine, Department of Neurology

227 E 30th St, 8th Floor, New York NY 10016 USA

[2] UC San Diego, Department of Psychology

9500 Gilman Dr., La Jolla CA 92093 USA

Author Note

*Corresponding author: Adam M. Morgan, NYU School of Medicine, Department of Neurology, 227 E 30th St., 8th Floor, New York NY 10016 USA. adam.morgan@nyulangone.org.

This research was supported in part by the National Institutes of Health under grant R01-HD051030 to Victor S. Ferreira and by the National Science Foundation under grant DGE-1144086 to Adam M. Morgan.

Declarations of interest: None.

Abstract

Syntax famously consists of abstract hierarchical representations, essentially instructions for combining words into larger units like sentences. Less famously, most theories of syntax also assume a higher level of abstract representation. Representations at this level comprise instructions for creating the hierarchical representations used to create sentences. To date, however there is no experimental evidence for this additional level of abstraction. Here, we explain why the existence of such representations would imply that, under certain circumstances, speakers should be able to produce structures they have never been exposed to, and we test this prediction directly. We ask: Given the right type of input, can speakers learn a syntactic structure without direct exposure? In particular, different types of relative clauses have different surface word orders. These may be represented in two ways: with many individual representations or one general representation. If the latter, then learning one type of relative clause amounts to learning all types. We teach participants a novel grammar for only some relative clause types (e.g., just subject relative clauses) and test their knowledge of other types (e.g., object relative clauses). Across experiments, participants consistently produced untrained types, implicating the existence of this higher level of abstract syntactic knowledge.

Keywords: artificial language learning, generalization, relative clauses, syntax

Language learners produce novel relative clause types without exposure: Evidence for a single, general syntactic representation of long-distance dependencies

Introduction

Puppies often grow up side-by-side with human infants, living in the same houses, eating off the same floors, and hearing the same language. In spite of this, the outcomes of their cognitive development – particularly with regard to language – are radically different. Within a few years, and with little explicit teaching (or possibly none at all; Ochs & Schieffelin, 1994), children learn tens of thousands of words. Puppies and other non-human animals, however, do not learn more than a few hundred words, with the notable exception of Chaser, a Border Collie who knew 1,022 words (Pilley & Hinzmann, 2013).

Words are not the extent of learning: children also come to know the patterns describing how words combine such that *'People eat tomatoes'* describes the world, *'Tomatoes eat people'* describes the plot of a campy 1970s horror movie, and *'Eat people tomatoes'* is just confusing. There is little evidence that non-human animals ever come to reliably use *syntax* in this way. What evidence does exist reveals behaviors that, compared to human abilities, are relatively limited, as in the case of an African gray parrot named Alex who was able to form phrases like "blue peg" (Kako, 1999; Pepperberg, 1981) and a bonobo named Kanzi who followed commands like "Take the tomato to the colony room" (Savage-Rumbaugh et al., 1993; Truswell, 2017).

But even interpreting Chaser's and Alex's abilities generously, Chaser was still only capable of word learning, and Alex only of the simplest of combinatorics. These pale in comparison to the kinds of linguistic abilities that humans – even very young ones – know. To be sure, human adult grammatical competence includes words and simple grammatical patterns. For example, determiners (e.g., "the") and nouns ("dog") combine to form noun phrases ("the dog"), and verbs and noun phrases combine to form verb phrases ("feed the dog"). But human language also includes non-contiguous units, or *long-distance dependencies*.

Take for example the relative clause in "the dog [that I fed __]." "The dog" is the object of "fed," but unlike in typical verb phrases, it does not appear in its canonical position just after the verb. In fact, it is not even in the same clause. The empty position where objects canonically appear is referred to as a *gap* (and represented with an underscore throughout). Despite the distance between the object and the gap, and without any explicit knowledge of dependencies or verb phrases, speakers of English effortlessly interpret "the dog" as the object of "fed."

Long-distance dependencies like this come in all kinds of varieties, including relative clauses, questions ("Which movie did you want to see __"), *tough*-constructions ("The news was tough to swallow __"), and topicalization ("I could do without whitefish, but lox I love __"). Furthermore, each of these varieties consists of a family of related structures, as in the following relative clauses:

(1) a. the dog [that __ bit me] *Subject Relative Clause*

b. the dog [that I fed __] *Direct Object Relative Clause*

c. the dog [that I threw the ball to __] *Indirect Object Relative Clause*

Each relative clause in (1) is named for the position of the *gap* (or, more precisely, the syntactic role that the modified noun is interpreted as having in the relative clause). Thus, the subject relative clause (SRC) has a gap in subject position, the direct object relative clause (DORC) has a gap in direct object position, and the indirect object relative clause (IORC) has a gap in indirect object position. These relative clauses have different surface word orders: While SRCs like the one in (1a) have *'that,' Verb, Direct Object* order, DORCs have *'that,' Subject, Verb* order and IORCs have *'that,' Subject, Verb, Direct Object, Preposition* order.

Thus, relative clauses form a family of structures with similar properties. For each type of syntactic role in a language – subject, direct object, indirect object, oblique object, etc. – the language may have a corresponding relative clause structure. The objective of the experiments presented here is to assess the representational underpinnings of these kinds of long-distance dependencies. Broadly speaking, there are two possibilities.

The first is that each of these structures could be represented separately, at least as initially learned. Given that even very young learners of English have likely been exposed to each of the types in their language, productive use of the various types may be learned from these separately-experienced structures, leading to distinct representations,[1] as illustrated in (2).

(2) *Multiple, distinct relative clause representations:*[2]

a.	Head Noun$_i$	[*that*	$\_\__i$	Verb	(Direct Object)	(...)]		SRC
b.	Head Noun$_i$	[*that*	Subject	Verb	$\_\__i$	(...)]		DORC

We call this the *Multiple Representations Hypothesis*.

The Multiple Representations Hypothesis is the dominant view among proponents of surface-oriented syntactic theories, most notably Constructionist approaches (Birner & Ward, 1998; Fillmore, Kay, & O'Connor, 1988; Goldberg, 1995, 2013). These theories hold that syntactic representations map more or less directly onto strings of words. They tend to eschew multiple ("deep") levels of representation, derivations, transformations, invisible functional elements ("empty categories"), and any other form of representation that is not directly observable in behavior. For instance, while many prominent theories hold that structural alternates such as passives and actives derive from a single underlying representation, surface-oriented theories view different surface forms as reflecting distinct underlying representations (Duffield & Michaelis, 2011; Goldberg, 2002; Gries & Stefanowitsch, 2004; Langacker, 2009). These theories are robust to a number of observations that the traditional picture of a purely abstract system struggles to accommodate. For example, to account for idiosyncratic argument structures – why one must use a preposition with the verb *dine*, as in "dine *on* steak," but not with the semantically similar verbs *eat* or *devour* – some item-based knowledge is clearly required.

[1] We use the term *representation* to refer to any type of syntactic knowledge ranging from, e.g., knowledge of agreement morphology to full sentence-length "tree" structures. Where various theoretical frameworks invoke distinctions between "rules," "schemas," "structures," "frames," "constructions," etc., we remain agnostic, as our main theoretical question does not hinge on the existence of any such categories.

[2] Here and in Example (3), parentheses denote optional arguments, and noun phrases are indicated by their corresponding syntactic roles (subject, direct object). Note that these sample representations are not intended to be formally adequate, but instead to give a general, theory-neutral caricature of each hypothesis.

Proponents of *experience-based* (or *usage-based* or *item-based*) theories of language acquisition also tend to espouse the Multiple Representations Hypothesis. *Experience-based* theories posit that learning is a process of first memorizing particular *items* (or *exemplars*), and then abstracting over these to arrive at a system of syntactic representations (Christiansen & Chater, 2015; Diessel, 2004; Diessel, Dąbrowska, & Divjak, 2015; Lieven, Behrens, Speares, & Tomasello, 2003; McCauley & Christiansen, 2011; Tomasello, 2000a, 2009). Such models are consistent with evidence suggesting that a significant portion of linguistic knowledge reflects direct experience (Bolinger, 1977; Duffield & Michaelis, 2011; Goldberg, 2002; Tomasello, 2000b). The resulting representations are surface-oriented. As a result, there is considerable overlap between experience-based acquisition theories and surface-oriented syntactic theories.

The second possibility is that, at some level of abstraction, relative clauses might all be represented as just one structure. That is, instead of, or perhaps in addition to, the structure-specific representations in (2), there might exist a single, general representation that accounts for all of the different surface types of relative clauses. Such a unitary representation might take the form of a general underlying principle like, *Clauses can modify nouns, but when doing so, the modified noun should be omitted from the clause.* If we were to sketch this representation, it might look like:

(3) Head Noun [*that* Subject Verb (Direct Object) (...)]**

**DO NOT REPEAT THE HEAD NOUN[3]

We call this the *Single Representation Hypothesis.*

While a single representation like Example (3) may seem unnecessarily baroque, this is in fact the dominant approach in theoretical syntax. There are many different approaches to formalizing such a representation, including *movement* (Government and Binding, Minimalism: Chomsky, 1981, 1992), *slash categories* (Head-Driven Phrase Structure Grammar, Simpler Syntax: Culicover & Jackendoff, 2005; Pollard & Sag, 1994), *functional uncertainty* (Lexical Functional Grammar: Bresnan, Asudeh,

[3] The specifics of the asterisked clause may vary depending on the particular relativization strategy of a language. For instance, in pronouns which use resumptive pronouns rather than gaps, it may be something like: **REPLACE THE REPEATED HEAD NOUN WITH A PRONOUN.

Toivonen, & Wechsler, 2015; Kaplan & Zaenen, 1989), *function composition* (Categorial Grammar: Steedman, 1993), and *feature percolation* (Sign-Based Construction Grammar: Sag, 2010). However, despite this preponderance of theoretical modeling, to our knowledge there exists no direct experimental evidence that these types of structures are psychologically represented. (See McElree and Bever (1989) for a related discussion on the psychological reality of gaps.)

If a single, general representation does exist, it has important implications for the taxonomy of abstract structures, and consequently for theories of syntax. It would require that there be an additional level of abstraction above simple hierarchical structures like those in (2). That is, if we think of the representations in (2) as blueprints for forming strings of words like those in (1), then the representation in (3) can be thought of as a blueprint for how to form other blueprints, like those in (2). Such a representation cannot be captured by a strictly surface-oriented theory of syntax, nor by acquisition theories which predict surface-oriented syntax.

Here we take advantage of the fact that the Single Representation Hypothesis and the Multiple Representations Hypothesis predict different patterns of learning behavior. Specifically, if the different surface forms of relative clauses each have distinct representations, it would predict that each type of relative clause must be learned individually. But if the different surface forms of relative clauses have a single underlying representation, then when one relative clause is learned, it is not learned as an isolate. Instead, others may "come along for the ride" because what is learned is a general representation for how a noun is modified with a clause. Such a finding would not only have implications for theories of syntax, but also for experience-based theories of language acquisition, which cannot readily accommodate the acquisition of a syntactic representation for which there are no exemplars.

We present four experiments in which adults learn an artificial language via exposure to sentences with relative clauses. These sentences, however, only contain a subset of the types of relative clauses given in (1). After training, participants are asked to use the new language to describe pictures, some of which elicit descriptions using the *trained* relative clauses – those which participants were exposed to – and others of which elicit *untrained* ones – those which participants were not exposed to. Generalization from trained to untrained structures is assessed to infer the nature of the underlying representation of long-distance dependencies.

If each type of relative clause has a distinct representation, then learning one structure does not implicate knowledge of other types, and we should observe little or no generalization. But if the various types of relative clauses have a single underlying representation, then learning one type may be equivalent to learning all types. On this account, we should expect to see that participants *do* generalize to untrained structures.

The remainder of the paper is structured as follows: In the following section, we give more detail about relative clauses, particularly from a typological perspective. We then discuss previous work validating the use of artificial language learning paradigms for studying syntactic representation. After presenting four experiments which test whether learners' knowledge of relative clauses is general or specific, we interpret findings and discuss implications for theories of syntax and language learning.

Relative clauses

There is a great deal of variation in the surface structure of relative clauses across languages. For instance, as (4) shows, English relative clauses (in brackets) appear after the noun they modify, the *head noun* (in bold). Relative clauses like this are called *postnominal.*

(4) the **girl** [that _ hugged the bear] POSTNOMINAL RELATIVE CLAUSE

But in many languages, relative clauses *precede* the head noun. Such *prenominal* relative clauses, as in (5), are the default for at least 141 documented languages including Amharic, Basque, Chinese, Huallaga Quechua, Malayalam, and Turkish (Dryer, 2013).

(5) a. *Malayalam prenominal relative clause*[4]

[4] Thanks to Savithry Namboodiripad and Sin Hang Lau for these examples.

[_ kaɾa dije ke tːipːi dit͡ɕːa] ku tːi
GAP bear.ACCUSATIVE hugged.COMPLEMENTIZER girl

'the girl who hugged the bear'

b. *Mandarin prenominal relative clause*

nàge [_ bào zhe xiǎo xióng de]
DEMONSTRATIVE GAP hug TENSE ADJECTIVE bear COMPLEMENTIZER
nǚhái
girl

'that girl who is hugging the bear'

If English were such a language, we might express the meaning of (4) and (5) with something like:

(6) the [that _ hugged the bear] **girl** PRENOMINAL RELATIVE CLAUSE

Another source of variation in relative clauses, mentioned earlier, depends on what role the head noun plays inside the relative clause. Examples (4) and (6) are both SRCs because the head noun, *girl*, is the subject of the relative clause verb *hugged*. Inside the relative clause, the subject role is not repeated, leaving behind a *gap*.

In principle, the gap can have any syntactic role that nouns can have. For instance, it could be in a subject, direct object, indirect object, oblique, etc. This is true for both post- and pre-nominal relative clauses. Table 1 schematizes these various types of relative clauses, with mock-ups of prenominal versions using English words.

In each of the four experiments presented here, participants are presented with a grammar similar to that in the right column of Table 1. That is, participants are told that they will be learning a language that is a composite of English words and Chinese grammar (Experiments 1a and 1b) or Korean (Experiments 2 and 3). They are trained on one or two types of prenominal relative clauses (for instance, only SRCs) and then tested on their knowledge of the trained type (SRCs) and an untrained type (e.g., DORCs).

Language learning as a tool to study syntactic representation

Our objective is to use relative clause learning as a tool to probe the nature of syntactic knowledge. We are not the first to approach language questions this way.

Table 1

Relative clauses vary along two dimensions: position of the relative clause relative to the head noun (columns) and syntactic role of the head noun inside the relative clause (rows).

Gap Position	Relative Clause Position Postnominal (English-type)	Prenominal (Chinese-type)
Subject	the **girl** [that _ hugged the bear]	the [that _ hugged the bear] **girl**
Direct Object	the **bear** [that the girl hugged _]	the [that the girl hugged _] **bear**
Indirect Object	the **bear** [that the girl gave porridge to _]	the [that the girl gave porridge to _] **bear**
Oblique	the **bowl** [that the girl served the porridge in _]	the [that the girl served the porridge in _] **bowl**
Object of Comparison	the **girl** [that the bear was furrier than _]	the [that the bear was furrier than _] **girl**

Previous studies have used similar paradigms to investigate topics such as *regularization*, a phenomenon whereby people impose grammatical structure on unstructured input (Culbertson & Smolensky, 2012; Culbertson, Smolensky, & Legendre, 2012; Kam & Newport, 2009; Saldana, Smith, Kirby, & Culbertson, 2018; see also Senghas, Kita, & Özyürek, 2004 for a longitudinal study of how this happened in a new natural language over several generations). Others have tested theories about specific pressures on language processing by teaching participants a grammar with, for instance, inefficient properties and showing that learners change the grammar to make it more efficient (Fedzechkina, Jaeger, & Newport, 2012). (See also Bley-Vroman, Felix, & Loup, 1988; Christiansen, 2001; Culbertson & Newport, 2017; Gass & Ard, 1980; Tily, Frank, & Jaeger, 2011.)

In the present study, we aim to measure *generalization.* Generalization has long been used as a tool to study representation – in particular, to study the abstractness of representations. For instance, Berko (1958) showed that children as young as 4 years who are introduced to a new word, such as *wug*, can correctly generalize to a previously unseen form of that word, *wugs.* In an artificial language learning paradigm, Culbertson and Adger (2014) exposed native English speakers to noun-adjective pairs like "horse big" and "houses two," following which participants were asked to combine a noun with two adjectives. Their participants produced utterances like "boxes big two," generalizing to previously unseen combinations that were inconsistent with the linear order of adjectives in their native language ("two big"), but consistent with the universally attested hierarchical order, thus experimentally demonstrating a bias toward hierarchical, rather than linear, organization of words. Similarly, Kaschak (2006) exposed adult speakers of Standard American English to the nonstandard *needs* construction, as in "the car needs washed." After a training regimen involving exposure to several instances of this construction using the verb *needs*, Kaschak showed that participants subsequently generalized to novel instances of the construction involving different verbs (e.g., "the dog *wants* walked") and different global syntax (a pseudocleft construction, e.g., "what the car needs is washed").

Four relevant points can be gleaned from this body of work. First, at least some syntactic representations are abstract. This accounts for the productivity observed in studies like Berko (1958) and Kaschak (2006). Second, at least some syntactic representations are hierarchical, not linear, accounting for participants' generalizations to word orders that are consistent with universal patterns (Culbertson & Adger, 2014; Saldana, Oseki, & Culbertson, 2019). Third, artificial language learning tasks present a valid way to approach questions about syntactic representation (Culbertson, 2012; Fedzechkina, Newport, & Jaeger, 2016; Tily et al., 2011). Finally, findings from artificial language learning tasks reveal behaviors that cannot be traced back to pre-existing knowledge of languages that participants already speak, and therefore reflect properties of human linguistic abilities in general (Culbertson, 2012; Culbertson & Adger, 2014; Saldana et al., 2019; Tily et al., 2011).

The present study

Following these approaches, we present a series of experiments in which monolingual English speakers are trained on artificial languages. We aimed to measure generalization in order to better understand the makeup of the representation of syntactic structures. Of particular interest is which of two possible representational systems underlies speakers' knowledge of long-distance dependencies: multiple specific representations or a single, general one.

In all four experiments, participants were taught artificial languages with prenominal relative clauses via exposure. English words were used so as to minimize the amount of training necessary. The languages differed from English in their word orders and morphological properties, which were systematically varied across experiments to parcel out the contributions of participants' knowledge of English.

Different groups of participants were trained on different subsets of the artificial languages. Some groups were trained only on prenominal SRCs, and others only on prenominal DORCs. After training, all groups were tested on both the trained and untrained structures. If each type of relative clause is learned independently and therefore has a distinct mental representation, as in the Multiple Representations Hypothesis, then participants should produce few or no untrained structures. But if upon being exposed to one type of relative clause, participants acquire a representation of relative clauses in general, as in the Single Representation Hypothesis, they should be able to produce relative clauses even of types they have not previously seen.

It should be noted that the type of generalization we investigate here is different from that in Berko (1958) and Kaschak (2006). In those studies, generalization consisted of producing a new *token* (e.g., *wugs*). To do so, participants used a grammatical representation that they already knew (e.g., "add -*s* for plural"). In the present study, we ask not whether people are able to generalize to new tokens, but to new *grammatical representations* – that is, to a syntactic representation of the untrained type of relative clause. The notion of generalization here is therefore more similar to that in Culbertson and Adger (2014) and Saldana et al. (2019).

Furthermore, the type of generalization we assess is also qualitatively different from other types of lexically-based generalizations, such as generalizing from a noun used to describe a tool ('the hammer') to a verb to describe the action of using that tool ('to hammer'). For the family of relative clause structures, a single 'meta-rule' like (3) would describe how all types of relative clauses are built out of constituent structures. In the lexical domain, there is no such ‘meta-rule’ that describes how, given one type of productivity (generalizing from ’the hammer’ to ’to hammer’), another type of productivity is licensed.

We controlled for a number of potential confounds. First, adult language learning and child language learning may be different, as there are different sources of information available to adults and infants (Bley-Vroman, 1989). Infants have whatever learning mechanisms are in place early in life (which may or may not be available to adults), while adults have adult reasoning abilities and whatever languages they already know. If adults use either of these sources of information, it may result in representations that are different from those we aim to understand – that is, those acquired by children in naturalistic settings.

For instance, adult language learning tends to be more veridical: whereas children will almost always impose structure to regularize an unstructured input, adults sometimes will (Culbertson & Smolensky, 2012; Culbertson et al., 2012; Fedzechkina et al., 2012), and other times will not (Culbertson & Newport, 2017; Kam & Newport, 2009). This may pose a problem for interpreting the results if we find that learners do not generalize to untrained structures. But if participants do generalize, then they do so despite a tendency to learn more veridically.

Another signature of adult language acquisition is the presence of transfer effects, where properties of the adult’s native language are observable in their nonnative language. This indicates that prior linguistic knowledge can impact the acquisition of a new system (Gass, 1979; Martin, Abels, Adger, & Culbertson, 2019). One concern, then, is that participants in this study – all monolingual English speakers – might generalize to untrained structures using their knowledge of English syntax rather than some newly-acquired representation.

However, such a possibility would still support the Single Representation Hypothesis. This is because the production of an untrained structure implicates the existence of a single, general representation of relative clauses, regardless of where that representation came from. Participants might learn a general representation during the training regimen in the present experiments and use that to produce untrained relative clauses, or they may simply rely on a general representation of English relative clauses. In either case, their production of untrained types of relative clauses implicates a single, general representation. In fact, the use of English knowledge may constitute even stronger support given the more ecologically valid nature of their English syntactic knowledge.

Another potential issue we aimed to avoid was participants developing and using explicit knowledge of the grammar. Syntactic knowledge is implicit knowledge, and so any production that is guided by explicit knowledge probably does not stand to shed light on the nature of syntactic representations. For instance, an adult might develop an explicit strategy for forming relative clauses, such as "produce a determiner, then a relative clause with the verb at the end, then the head noun." If participants used such a strategy, they may have been able to produce untrained structures without actually generating a syntactic representation.

However, explicit descriptions of complex structures like relative clauses are extremely difficult to formulate. Anecdotally, it often takes several semesters of syntax coursework for undergraduate students to understand the syntax of long-distance dependencies. As such, this was regarded as a relatively minor concern. We nonetheless took steps to determine whether participants used explicit knowledge. First, we included an "EXPLICIT" training group in Experiment 1a, in which participants received a grammar lesson on prenominal relative clauses, including seeing written examples and descriptions of gaps and the function of relative clauses. No other participants in any of our experiments were presented with written stimuli, nor did they receive any explicit description of gaps, relative clauses, etc. Experimenters were instructed to never make metalinguistic comments when correcting participants (e.g., "don't say the word *that*" or "put *hugged* before *bear*"), but instead to simply repeat the correct response in full so as not to encourage the development of explicit strategies (see Exp. 1a Procedure).

If explicit knowledge underlies the behavior of the implicitly trained groups, then we should expect the implicit and explicit groups to perform similarly. Specifically, if the implicit groups are able to correctly produce prenominal relative clauses at test and the EXPLICIT group is not, it suggests that the success of the implicitly trained groups cannot be attributed to the type of explicit knowledge the EXPLICIT group had – at least not exclusively. Second, after the experiment all participants completed a debriefing survey. Participants were required to respond to the prompt, “Describe the sentences we trained you on. Can you explain the rules to follow to make them?” We excluded data from any participants who provided a response that coherently articulated the grammar.

We also intentionally designed the grammars to be different from English. Specifically, in all experiments the grammars contained finite prenominal relative clauses, which do not exist in English (although see Eilish, 2019). While the grammar of Experiment 1 was otherwise similar to English, the grammar of Experiment 2 had a different basic word order, and Experiment 3 included both a different basic word order and case marking on nouns. These differences mean that the type of explicit rules that a participant would have to formulate to correctly produce trained and untrained sentences would be highly complex, and increasingly so with each new experiment. Thus, it was decreasingly likely that participants would be able to formulate explicit grammatical knowledge.

A final consideration we kept in mind when designing stimuli had to do with typological validity. When considering grammatical properties across languages, a number of distributional patterns emerge. There is a divide in the field between researchers who view these as the result of processing pressures or language contact effects (Dunn, Greenhill, Levinson, & Gray, 2011; Evans & Levinson, 2009; Fitz, Chang, & Christiansen, 2011; MacDonald, 2013) and those who think that they might reflect more rigid constraints on language (Bley-Vroman et al., 1988; Culbertson & Adger, 2014; Gass, 1979; Gass & Ard, 1980). If the latter, then grammars which are unattested may in fact be impossible to learn with the usual language processing architecture. Any syntactic representations acquired from such an impossible grammar might be different from native-like representations in some important way, and this could limit the interpretability of our findings. We do not take a position on this debate, but out of an abundance of caution we attempted to design our stimuli so as to conform to typologically attested grammars, particularly in Experiments 2 and 3.

Experiment 1a

In Experiment 1a, participants were trained on an artificial language with prenominal relative clauses. After training, their knowledge of the grammar was tested in a production task where they described pictures using the new grammar. Two groups were trained on only one type of relative clause: a SRC-ONLY group, trained only on SRCs, and a DORC-ONLY group, trained only on DORCs. These groups are collectively referred to as the ONLY groups. After training, both groups were tested on their knowledge of both prenominal SRCs and DORCs. If different types of relative clauses have distinct syntactic representations, then we expect participants not to generalize to the UNTRAINED structures. If, on the other hand, there is one general underlying representation for all types, then participants should be able to produce the UNTRAINED structures.

Previous research has demonstrated that learning can be improved by a more diversified input. This has been shown in the memory literature (*desirable difficulty* effects; Bjork, 1994), and recently in lexical learning and processing (Hoffman, Ralph, & Rogers, 2013; Johns, Dye, & Jones, 2016; Johns, Gruenenfelder, Pisoni, & Jones, 2012; Plummer, Perea, & Rayner, 2014). On the hypothesis that contextual diversity might similarly boost acquisition of syntactic structures, a third "BOTH" group received exposure to SRCs and DORCs. This group was tested on two TRAINED structures (and no UNTRAINED ones).

To ensure that the ONLY groups learned implicit representations, not explicit ones, participants were trained with direct exposure to sentences rather than explicit grammar lessons. However the possibility remained that such training might allow participants to develop explicit knowledge of the structure. A fourth group was therefore trained with an explicit grammar lesson so as to compare the performance of the implicitly trained groups (SRC-ONLY, DORC-ONLY, and BOTH) to performance that unambiguously reflected explicit knowledge.

Experiment 1a therefore was designed to answer the following questions: (1) Do participants trained on only one type of structure generalize to the other type? (2) If so, can this be attributed to explicit grammatical knowledge? (3) Does learning of trained structures benefit from a more syntactically diverse input?

Method

Participants. Participants were continuously run until a target of 24 per group was met.[5] A total of 112 UC San Diego undergraduates participated for course credit. Pre-screen requirements were the same for all experiments: these included that participants were over 18 years old and were native monolingual speakers of English (defined as not having learned any language but English before the age of 7). A total of 16 participants were excluded: 2 for knowing a language with finite prenominal relative clauses (such as the ones in the artificial language), and 14 due to software or experimenter errors.

Factors. The experiment had one between-subjects factor with four levels, training GROUP, and one within-subjects factor with two levels, TRIAL TYPE. Participants were randomly assigned to one of four training groups. The SRC-ONLY group received implicit training on 36 prenominal SRCs; the DORC-ONLY group received implicit training on 36 prenominal DORCs; the BOTH group received implicit training on 18 prenominal SRCs and 18 prenominal DORCs; and the EXPLICIT group received an explicit grammar lesson on prenominal relative clauses but no implicit training (other than whatever may have been gleaned from a single example of a prenominal DORC). After training, all groups were tested on 36 previously unseen items. Items appeared in one of two conditions: either TRAINED trials, where the stimulus was designed to elicit a trained structure (SRCs for the SRC-ONLY group; DORCs for the DORC-ONLY group, and SRCs or and DORCs for the BOTH GROUP) or UNTRAINED

[5] Our lab's standard sample size for experiments with within-subjects manipulations is 48. We doubled this due to the between-subjects group manipulation, resulting in 24 participants per group.

trials (DORCs for the SRC-ONLY group; SRCs for the DORC-ONLY group, and SRCs and DORCs for the EXPLICIT group). This TRIAL TYPE manipulation was counterbalanced across items such that for a given item, half of participants in each group saw it in a TRAINED trial and half saw it in an UNTRAINED trial.

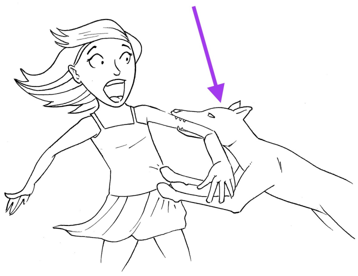
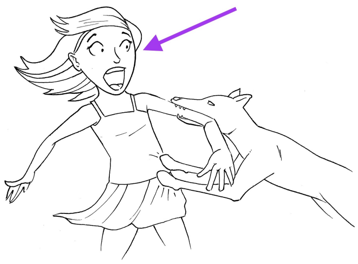
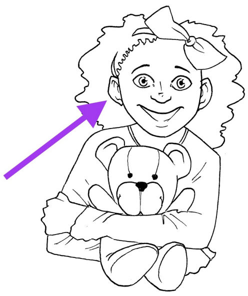
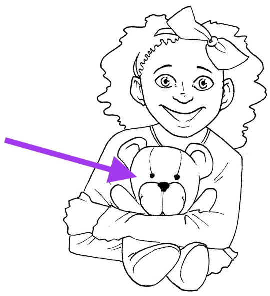
Materials. In all experiments, the artificial language used English words but non-English syntax. This greatly reduced the amount of training relative to a more typical task where both a new lexicon and grammar would have to be taught. The syntax of the relative clauses in Experiment 1a and 1b was identical to English except in that the relative clauses appeared prenominally rather than postnominally. Each sentence started with functional material that introduced a noun, for example, "Here's the. . . ," or "These are some. . . ," or "Now we have a. . . ." This was followed by a transitive relative clause headed by a clause-initial complementizer ("that") and either a subject gap or a direct object gap, as in "__ hugged a bear" or "the girl hugged __." The relative clause was in turn followed by the head noun, which for some items was a bare noun ("girl") and for others contained a prenominal adjective ("little girl").

It is worth noting that this grammar is odd from a typological perspective. Cross-linguistically, languages with prenominal relative clauses tend to put complementizers at the end of the clause, and while there may be a few exceptions to this (e.g., Amharic, Laz, and Tigre; see Wu, 2011 for a full discussion), to our knowledge none of these languages have clause-initial complementizers in prenominal relative clauses. We nonetheless used clause-initial complementizers because we worried that without them, DORCs would be too odd given that they would require two adjacent determiners, as in ". . . the [the girl hugged __] bear." Indeed, natural languages deal with this situation in peculiar ways, as in the case of St'át'imcets, a Northern Interior Salish language which simply deletes one of the two adjacent determiners (Davis, 2010). Concerns relating to the typological validity of complementizers are addressed in subsequent experiments, where they are removed altogether.

Sample materials appear in Table 2. Training materials consisted of 36 pairs of images of transitive events, each with an accompanying sentence using a prenominal relative clause to describe one of the objects in the picture. Each pair of images showed the same event, but in one, an arrow pointed to the subject, and in the other an arrow pointed to the direct object. For images with a subject arrow, the accompanying sentence used a prenominal SRC. For images with a direct object arrow, the accompanying sentence used a prenominal DORC.

Table 2

Sample materials from Experiments 1a and 1b.

	SRC	DORC
Train	 “That’s the [that __ bit the little girl] dog.”	 “That’s the [that a dog bit __] little girl.”
Test	 *hugged*	 *hugged*

Testing materials consisted of 36 additional pairs of images, similarly created in pairs with arrows pointing to subjects or objects. These images were paired with verbs which were presented on the screen underneath the image during the test phase.

Procedure. All participants began by providing informed consent. They were then instructed that they would be trained on a new language which used English words but Chinese word order.[6] The task took about 60 minutes for participants in the implicit training groups (SRC-ONLY, DORC-ONLY, and BOTH) and under 30 minutes for

[6] There are some differences between the word order of Chinese and that used in our stimuli. For instance, Chinese does not have relative clause-initial complementizers like the one we used. This was not explained to participants.

participants in the EXPLICIT group.

For participants in the implicit training groups, the experiment had three phases: two training and one test (see Figure 1). In the first training phase, participants saw a picture appear on a monitor with an arrow pointing to either the agent or patient (as determined by their group) and the experimenter read aloud the corresponding sentence. The participant was instructed that they would have to repeat the sentence soon, so to listen carefully and ask the experimenter to repeat as needed. After two trials like this, the two images from the preceding two trials appeared one at a time in random order with a prompt indicating that the participant should try to recall the sentence as the experimenter had said it. If the response was correct, they went on to the next trial. If it was incorrect, the experimenter corrected them.

Responses during training phases were considered correct if they contained a grammatically well-formed prenominal relative clause with an overt complementizer (“that”) and the correct meaning. This meant that the specific words could differ from those that the experimenter had initially used to describe the picture; lexical substitutions like “nurse” for “doctor” or “the” for “a” were not considered errors. Experimenters, who were undergraduate RAs, were told that if they were ever in doubt about whether a participant’s utterance was correct, to read the sentence again and ask the participant to repeat it verbatim.

Experimenters were instructed not to correct participants until they had made a reasonable attempt at recall (following findings from the learning literature that learning is enhanced when learners are tested rather than given information during training, e.g., Kang, Gollan, & Pashler, 2013; see Roediger & Karpicke, 2006 for a review). Experimenters did not give corrections in the form of meta-linguistic commentary (e.g., “remember to say ‘that’” or “use past tense”), which we worried might facilitate explicit learning, but to simply repeat the full sentence if the participant made any errors. If the participant was not able to say it correctly after three tries, then the experimenter could go on to the next trial so as to minimize frustration. The first training phase ended after all 36 training images had been presented and recalled in this fashion.

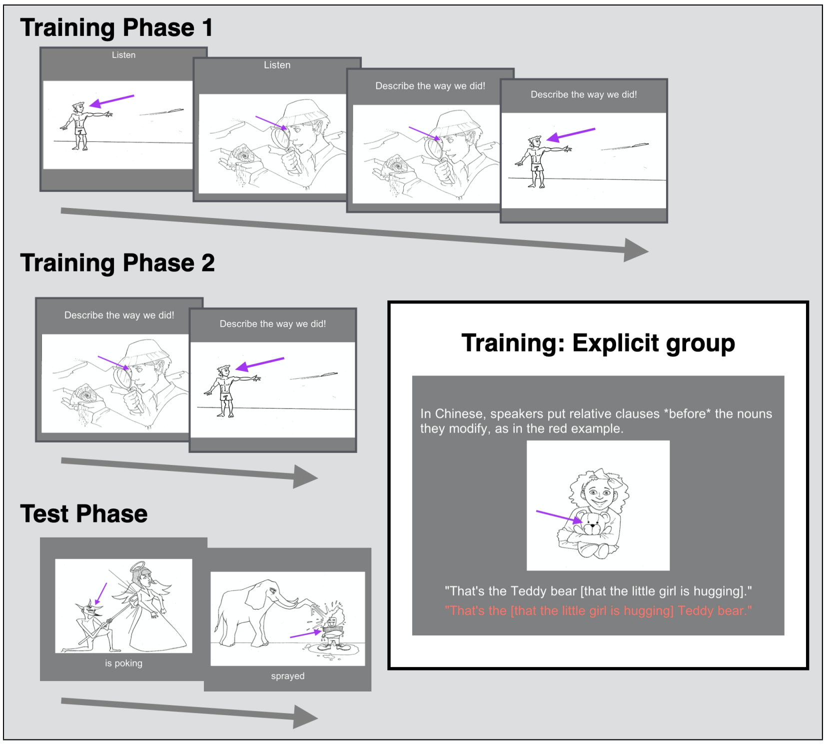


Figure 1. Experiment 1 procedure. In Training Phase 1, participants in the SRC-ONLY, DORC-ONLY, and BOTH groups saw 36 pictures, twice each, in pairs. The first time they saw a picture, the experimenter described it aloud with a prenominal relative clause, e.g., “That’s the that threw a frisbee guy.” After two pictures were described like this, the same two pictures appeared in random order and the participant was asked to repeat the description. In Training Phase 2, the same 36 images appeared in random order. Participants described each one as they had in Phase 1. The EXPLICIT group had only one training phase in which participants heard only one example of a prenominal relative clause. All four groups received the exact same stimuli at test: 18 previously-unseen images eliciting SRCs and 18 eliciting DORCs.

In the second training phase, participants saw the same 36 training images again, presented one at a time in random order and with the verb written underneath. They were instructed to try to recall the sentences as they had in the previous phase. Again, experimenters corrected them as needed by reading aloud the full sentence, but only after they had attempted to recall the sentence.

Instead of 36 example sentences, the EXPLICIT training group received a formal grammar lesson accompanied by a single example of one prenominal DORC and its postnominal translation (as in Table 1). Experimenters let participants read these examples on the screen and then gave a more detailed explanation. They explained that relative clauses are sentences, like "The girl hugged the bear," which modify nouns, like "girl" or "bear." Using the on-screen postnominal (English-type) relative clause, they pointed out that the noun being modified is not repeated inside the relative clause. They then explained that while in English, relative clauses appear after the noun they modify, in other languages like Chinese, the relative clause appears before the modified noun. During the course of these instructions, they read the prenominal relative clause aloud exactly once. Participants were instructed to ask as many questions as they wanted before they would be asked to produce prenominal relative clauses on their own without receiving help.

After training, all groups went on to the same test phase. Participants were instructed that they would describe 36 brand new images using their newly acquired grammar. They were told that the experimenters could not provide feedback during this phase, but may occasionally ask them to repeat the full sentence fluently to facilitate later transcription of the audio recording. All participants in the implicit training groups saw 18 images that elicited SRCs and 18 that elicited DORCs. For the BOTH and EXPLICIT groups, these appeared in random order, while for the SRC-ONLY and DORC-ONLY groups the order was pseudo-randomized such that the first four trials always elicited TRAINED structures. This was intended to facilitate transition to the test phase before (implicitly) asking them to generalize to the untrained structure.

Data coding and analysis. After the experimental session, research assistants manually transcribed and coded responses. Data from 22 trials were excluded because the trial was inadvertently skipped, a relative clause was produced with an intransitive verb,[7] or an incorrect response was due to experimenter error (e.g., if the experimenter did not correct a misinterpretation of the depicted event). The remaining responses

[7] Fox (1987) has argued that intransitive subjects may be more similar to direct objects than to transitive subjects when it comes to relativization from the perspective of processing and typology; we therefore excluded the few trials where participants produced intransitive active-voice SRCs so as not to inadvertently give one group an advantage.

were coded for structure type and for what errors they contained (if any). Structure types could be SRC, DORC, or uncodable (any utterance that could not clearly be labeled as a SRC or a DORC or that contained three or more errors).

As is not uncommon in studies with sentence production DVs, there was considerable variability in the responses (exuberant responding; Bock, 1996). To simplify matters and avoid potentially introducing experimenter bias, we adopted stringent criteria for coding responses as correct or incorrect (our dependent variable). Responses were coded as correct if they were grammatically well-formed instances of the elicited structure and if they conveyed the correct meaning. To be considered grammatically correct, the particular words did not matter (i.e., lexical substitutions were allowed, like *woman* for *girl*), but the response had to be a full sentence with a prenominal finite transitive relative clause. Any other errors meant that the trial was coded as incorrect, with one exception. Due to the exceptionally high number of instances where a relative pronoun (e.g., “who”) was used rather than the overt complementizer (“that”), or where neither was used, we counted all three of these alternatives as correct in all analyses. This error and the other most common errors were labeled and their prevalence is given in Table 3.

For all experiments, logistic mixed effects regressions were used to model responses as a function of GROUP and TRIAL TYPE (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2018); both factors were treatment coded. All models contained random intercepts for participants and items, and all fixed effects were allowed to vary by random factors if the effect varied within the factor. Following Barr, Levy, Scheepers, and Tily (2013), we report the maximal random effects structure which allowed the model to converge: random effects correlations and then random slopes were removed from the full model one by one in order from least variance accounted for to most until the model converged. For all analyses, we report the model output, and for significant effects of theoretical interest, we also report the results of model comparisons to determine whether each fixed effect contributed significantly to model fit.

For the production of untrained structures to be evidence in favor of the Single

Table 3

Percentage of the most common errors in Experiment 1a, by condition.

Error	BOTH Trained	EXPLICIT Untr.	SRC-ONLY Tr.	SRC-ONLY Untr.	DORC-ONLY Tr.	DORC-ONLY Untr.	Overall
Missing head determiner	11.92	0.81	15.74	15.04	21.76	18.52	12.07
Relative pronoun*	10.07	2.08	11.11	12.50	11.34	12.27	8.94
English	0.69	29.63	0.93	1.16	4.17	3.93	8.85
Wrong type	4.05	1.85	0	28.7	1.39	1.13	6.65
Missing determiner in RC	4.75	1.04	2.08	0.93	6.94	8.10	3.70
Number agreement error	1.74	1.97	1.39	3.47	9.95	6.02	3.57
No complementizer*	2.55	5.32	3.70	1.39	3.24	2.78	3.36
Head determiner after RC	0.69	1.50	4.17	3.70	4.63	4.63	2.69
Repeated head determiner	1.85	4.05	1.85	1.39	2.31	3.24	2.58

* While these were different from the input grammar, they are counted as correct in all Exp. 1a analyses.
Note 1. Errors were not mutually exclusive. Errors from responses that had three or more errors are not reported in this table as it was often too difficult to determine which particular errors led to the response.
Note 2. Example correct SRC response: "Here's the that hugged the bear girl." Example missing head determiner response: "Here's that hugged the bear girl." Example relative pronoun response: "Here's the who hugged the bear girl." Example English response: "Here's the girl that hugged the bear." Example missing determiner inside RC response: "Here's the that hugged bear girl." Example number agreement error: "Here's the that the grandma baked cookies." Example no complementizer: "Here's the hugged the bear girl." Example head determiner after RC: "Here's that hugged the bear the girl." Example repeated head determiner: "Here's the that hugged the bear the girl."

Representation Hypothesis, such productions must not be attributable to chance performance. We therefore aim to compare the rate of production of untrained structures to whatever rate would be expected by chance. However, "chance" is difficult to define for the present experiments. One approach might be to equate it with the likelihood of arriving at a correct structure by randomly ordering the words in an utterance. For a 7-word utterance like "Here's the that hugged the bear girl," any particular string has a $\frac{1}{7!} = 0.02\%$ chance of arising simply by chance. However, even participants who learned nothing about the grammar in the training phase would presumably be much less likely to place a determiner before a verb than a noun in a "random ordering" response strategy. Thus, 0.02% would be much too low a bar.

Rather than chance, we aimed to compare performance to what we call the *educated guessing rate.* We assumed that participants who produced fewer than 50% of the trained structures correctly at test were less likely to have acquired the structural representations necessary to produce the untrained structures. The rate at which these participants produced untrained structures could be therefore interpreted as an estimate of how often a participant would correctly produce untrained structures by making an educated guess. Note that this a particularly liberal estimate of the educated guessing rate given that some of the participants in this group may indeed have acquired the target syntactic representations.

The logic of the statistical test was therefore: If those participants who produced trained structures correctly on *more* than 50% of trials also produce untrained structures at a rate significantly higher than the educated guessing rate, then they were not simply guessing. Instead, they may have been engaging in syntax-driven production, implicating the existence of a general representation of relative clauses.[8] Note that this approach assigns no particular meaning to 50%, which is simply a convenient cutoff that, across experiments, splits participants into roughly even groups.

For each experiment we therefore report a second analysis where we separate participants into two subsets: those who produced more than 50% of the trained structures correctly and those who produced fewer than 50% correctly. These analyses are mixed-effects logistic regressions formulated as above. The dependent variable is whether the untrained structure is produced correctly. Fixed-effects terms were GROUP and SUBSET, the latter of which had levels $> 50\%$ and $< 50\%$ (i.e., whether participants produced the trained structure correctly on more or less than 50% of trials). As above, both factors were treatment coded.[9]

If production of untrained structures simply reflects guessing, then we should expect the amount of generalization between these two groups not to differ. However, if generalization reflects learning of a generalized representation for relative clauses, then

[8] To be clear, we should also expect participants to make *some* errors at test, and this should not be taken as evidence that they did not learn. The acquisition of a structure does not require that participants' responses are 100% accurate. (Consider adults who move to another language community and learn the local language: they can clearly be said to be learning, but also commit many errors along the way.)

[9] Thanks to Eva Wittenberg for suggesting this clever approach.

we should expect that the subset that learns the representation better will produce more of the untrained structures (i.e., a main effect of SUBSET).[10]

Critical predictions. For all experiments, our main theoretical question is whether speakers have a general representation of relative clauses, as in the Single Representation Hypothesis. Statistically, we ask whether participants correctly produce the untrained structure more often than what would be expected by guessing. This is addressed by the second analysis, where participants' performance on untrained trials is predicted by their accuracy on trained structures. For all experiments, the prediction is the same: If participants use a general representation of relative clauses to produce the untrained structure, then we should see that participants in the $> 50\%$ subset produce untrained structures more than participants in the $< 50\%$ subset.

Results

Data are shown in Figure 2 and model results appear in Table 4. Our first a priori questions was: Do the ONLY groups generalize to the untrained structures? That is, does the SRC-ONLY only group learn DORCs and does the DORC-ONLY learn SRCs? We addressed this with Model 1, a 2×2 model of response as a function of TRAINED vs. UNTRAINED structure and GROUP: SRC-ONLY and DORC-ONLY. Data from the BOTH and EXPLICIT groups were excluded from this analysis. This model converged with the full random effects structure without random effects correlations.

There was a main effect of GROUP (confirmed by model comparison; $\chi^2(1) = 6.588$, $p = .010$) whereby the DORC-ONLY group performed less well than the SRC-ONLY group. There was a main effect of TRIAL TYPE ($\chi^2(1) = 41.094$, $p < .001$), whereby participants were less correct on untrained trials than on trained trials. Finally, there was an interaction ($\chi^2(1) = 7.560$, $p = .006$) reflecting the fact that the DORC-ONLY group produced more untrained structures than the SRC-ONLY group.

To determine whether generalization was above what would be expected of

[10] An alternative approach, suggested by a reviewer, is to use each participant's mean accuracy on trained trials as a continuous predictor of their performance on untrained trials. Results of such analyses, reported for each experiment in Appendix A, pattern with the results of the subset analysis reported in the main body of the paper.

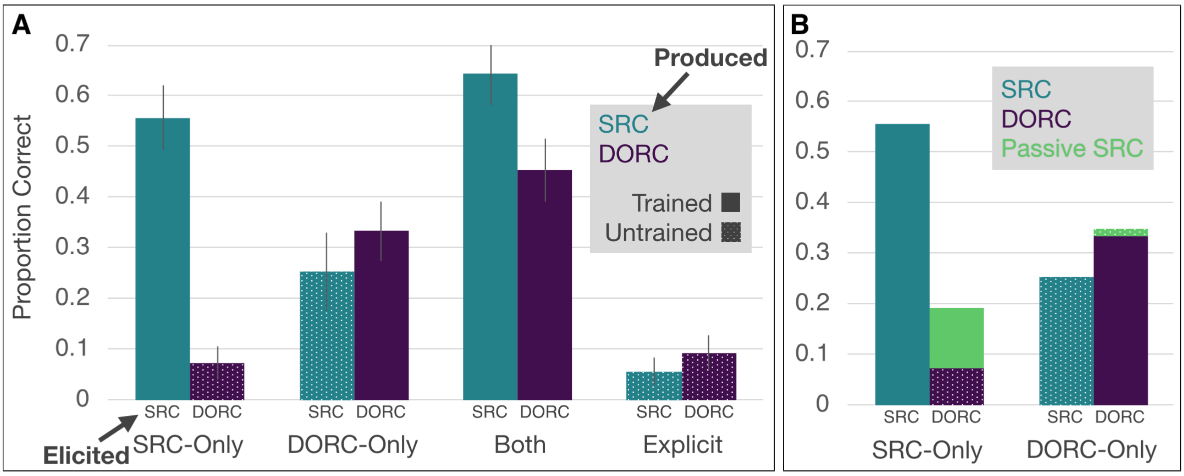


Figure 2. Experiment 1a results: (A) All correct responses and standard errors as a function of group and elicited structure (SRC or DORC). Colors indicate the type of structure produced and dots indicate untrained structures. (B) The ONLY groups again, but including well-formed passive SRC productions.

(educated) guessing, we looked for a relationship between learning of trained structures and generalization to untrained structures (Figure 3). We performed a subset analysis, comparing the amount of generalization by participants in ONLY groups who produced more than 50% of trained structures correctly to those who produced fewer than 50% correctly. The model converged with the full random effects structure. The results of this model, Model 1.2 in Table 4, show that the 21 higher-performing participants produced significantly more untrained structures (10% for the SRC-ONLY group and 67% for the DORC-ONLY group) than the 27 lower-performing participants (2% for the SRC-ONLY group and 11% for the DORC-ONLY group; model comparison confirmed that the effect contributes to model fit: $\chi^2(1) = 4.860$, $p = .027$). This indicates that generalization, at least among the high-performers, does not reflect guessing.

The results of Model 1 indicate that the SRC-ONLY group did not generalize to untrained structures as much as the DORC-ONLY group. But as Figure 2b shows, the SRC-ONLY group produced an unexpected number of passive structures in DORC-TRIAL trials. Passivizing a DORC results in a SRC with the same meaning as the original DORC. That is, the meaning of the DORC in “the bear [that the girl hugged __]” can be expressed just as well with the passive SRC, “the bear [that __ was hugged (by the

Table 4

Experiment 1a results.

	Model results			
	β	z	p	
Model 1: ONLY *groups* × TRIAL TYPE				
Intercept	0.276	0.785	.432	
GROUP: DORC-ONLY	−1.305	−2.626	.009	**
TRIAL TYPE: UNTRAINED	−4.957	−6.599	< .001	***
Interaction	2.874	3.103	< .001	***
Model 1.2: ONLY *groups'* UNTRAINED *trials, split by performance on* TRAINED				
Intercept	−10.182	−4.237	< .001	***
GROUP: DORC-ONLY	4.825	2.039	.041	*
SUBSET: >50%	5.719	2.402	.016	*
Interaction	0.623	0.220	.826	
Model 1.3: A re-run of Model 1, but with passive SRCs coded as correct				
SRC-ONLY, TRAINED (intercept)	0.280	0.780	.435	
DORC-ONLY, TRAINED	−1.231	−2.427	.015	*
SRC-ONLY, UNTRAINED	−2.453	−6.040	< .001	*
DORC-ONLY, UNTRAINED (interaction)	0.828	1.340	.180	
Model 2: Just untrained trials (EXPLICIT and ONLY groups)				
Intercept	−5.666	−5.850	< .001	***
GROUP: ONLY	2.543	2.274	.023	*
TRIAL TYPE: DORC	.0170	0.235	.814	
Interaction	−2.571	−1.983	.047	*
Model 3: Just trained trials (BOTH and ONLY groups)				
Intercept	0.904	2.562	.010	*
GROUP: ONLY	−0.621	−1.284	.199	
TRIAL TYPE: DORC	−1.228	−3.681	< .001	***
Interaction	−0.213	−0.331	.741	

girl)].” The SRC-ONLY participants discovered a way to respond correctly on DORC-eliciting trials without having to generalize to the untrained structure. Thus it is not clear whether the SRC-ONLY group did not generalize to DORCs because they did not know how, or because there was an easier strategy – passivization – available to them.

To determine whether the asymmetry in generalization indexed by the interaction term in Model 1 might have disappeared if SRC-ONLY participants had successfully produced DORCs on every trial in which they produced a passive SRC, we re-ran Model 1 on updated data in which all well-formed passive SRCs produced in

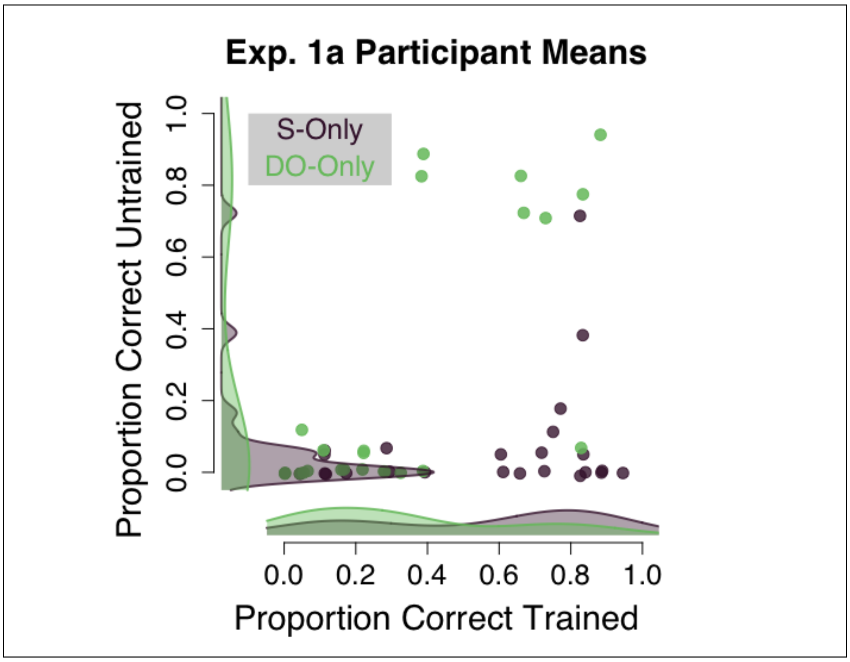


Figure 3. Experiment 1a: The relationship between learning trained structures (horizontal axis) and generalization to untrained structures (vertical). Individual participants (dots) are jittered with standard deviation of .005. Density plots appear along axes. Participants who learned the trained structures better also generalized more, suggesting that generalization did not reflect guessing.

DORC-TRIAL trials were coded as correct. This model, Model 1.3 in Table 4, converged with a full random effects structure but without random effects correlations. The results still show a clear difference between trained and untrained structures. However, the interaction representing asymmetrical generalization was no longer significant, meaning that there was no evidence that the DORC-ONLY participants were better at producing untrained structures than SRC-ONLY participants.

Our second question was whether generalization to untrained structures might reflect use of explicit knowledge of the grammar during production. This question was addressed by Model 2, which compared the ONLY groups' performance on untrained structures to the EXPLICIT group. Because there were no TRAINED trials in this analysis, the TRIAL TYPE factor here was coded as SRC-eliciting and DORC-eliciting. No data from the BOTH group nor the TRAINED structures for the ONLY groups were included. The model converged with the full random effects structure but without random effects correlations.

Results of Model 2 showed that the explicit group's production of SRCs and DORCs did not significantly differ (no main effect of TRIAL TYPE $\chi^2(1) = 0.052$, $p = .820$). The ONLY groups produced significantly more UNTRAINED structures than did the EXPLICIT group (a main effect of GROUP; $\chi^2(1) = 4.935$, $p = .026$). There was a significant interaction, reflecting the fact that the DORC-ONLY group generalized more than the SRC-ONLY group, however this should be interpreted with caution as model comparison revealed that it only marginally contributed to model fit ($\chi^2(1) = 3.674$, $p = .055$).

Our third question was whether participants' learning of trained structures would benefit from a more syntactically diverse input – specifically, if the BOTH group would learn SRCs better than the SRC-ONLY group and DORCs better than the DORC-ONLY group. Model 3 converged with the full random effects structure but without random effects correlations. Results did not provide evidence for the hypothesis: there was no main effect of GROUP ($\chi^2(1) = 1.629$, $p = .202$), although there was a main effect of TRIAL TYPE, reflecting the fact that, across groups, DORCs were learned less well than SRCs ($\chi^2(1) = 11.552$, $p < .001$). The interaction was also not significant ($\chi^2(1) = 0.107$, $p = .743$).

Discussion

Experiment 1a showed that learners are indeed capable of producing previously unseen structures: participants in the DORC-ONLY group generalized to untrained structures, although participants in the SRC-ONLY group appeared not to. This finding is supported by the fact that participants who learned trained structures better also generalized to untrained structures more, suggesting that generalization does not simply reflect guessing. Furthermore, given that EXPLICIT participants received explicit training on prenominal relative clauses, their performance may be viewed as a liberal estimate of how often one might guess correctly. The DORC-ONLY group performed significantly better than the EXPLICIT group, further supporting the idea that this group successfully generalized. The SRC-ONLY group, on the other hand, performed comparably to the EXPLICIT group on untrained structures.

The EXPLICIT group performed worse at test than any other group. There are many possible reasons for this. One possibility is that the explicit training regimen was not extensive enough for participants to acquire explicit knowledge of the prenominal relative clauses. Or perhaps participants *did* gain explicit knowledge, but this kind of knowledge is not easy to use to produce complex structures like relative clauses. An intriguing possibility, with potential implications for L2 pedagogy, is that explicit training is only beneficial in conjunction with implicit training. For instance, it might facilitate acquisition of implicit representations by allowing learners to practice producing and comprehending correct structures prior to their having acquired the implicit representation.

While the results of Model 3 did not show statistical evidence for a benefit associated with a syntactically diverse input, it is worth noting that the BOTH group performed numerically better on trained structures than either of the ONLY groups, and this was despite receiving only half the exposure to each trained structure. This suggests a possible benefit for diverse inputs that went undetected due to a lack of statistical power.

Finally, a post-hoc analysis (Model 1.3) revealed a potential explanation for why the SRC-ONLY group generalized less than the DORC-ONLY group. On several UNTRAINED trials, SRC-ONLY participants produced a passive SRC that both had the same meaning as the target DORC sentence and was of the TRAINED type (Figure 2b).

There are a number of possible reasons for this. For instance, it may have been easier for SRC-ONLY participants to produce passive subject relative clauses given that these were congruent with their training. If so, the lack of generalization could reflect participants opting for the more practiced subject relative clause structure rather than attempting the less-familiar direct object relative clause.

Another possibility is that participants produced passives due to processing pressures outside the scope of this investigation. Previous work on relative clause processing has revealed a tendency for speakers to produce passive subject relative clauses more often than active direct object relative clauses when the two arguments are both animate (as was the case for eleven of our 72 items) or when the active object is higher in animacy than the passive subject (true of five items) (Gennari & MacDonald, 2009; Gennari, Mirković, & MacDonald, 2012; Humphreys, Mirković, & Gennari, 2016; Roland, Dick, & Elman, 2007). It is therefore possible that these features of our stimuli contributed to the tendency for the SRC-ONLY group to produce passives.

For present purposes, however, what matters is whether SRC-ONLY participants were *able* to produce direct object relative clauses but did not because they could produce a passive structure instead (consistent with the Single Representation Hypothesis), or whether they were simply unable to produce direct object relative clauses (consistent with the Multiple Representations Hypothesis). Experiment 1b was designed to prevent participants from giving passive responses so as to encourage them to produce direct object relative clauses.

Experiment 1b

Experiment 1b was a partial replication of Experiment 1a. Two groups were trained on prenominal relative clauses: a SRC-ONLY group received training only on SRCs and a DORC-ONLY group received training on DORCs. The experiment differed from Experiment 1a in that the verbs given in the test phase were presented in unambiguously active-voice forms (e.g., “was hugging”) rather than in the simple past (e.g., “hugged”), which is often ambiguous between simple past and the participle form used in passives. Participants were instructed to use the verbs as written. If the asymmetry in generalization observed in Experiment 1a reflects SRC-ONLY participants not generalizing because they were not able to, then we should continue to see the asymmetry in this study. If, however, participants did not generalize simply because an easier response strategy was available to them, then we should no longer see an asymmetry in generalization.

Method

Participants. Participants were continuously run until the target of 24 per group was reached. A total of 52 UC San Diego undergraduates participated for course credit. Four participants were excluded: two for natively speaking a language other than English, one due to experimenter error, and one for producing a high number of passive SRCs during the test phase.

Factors. Two factors were manipulated: training GROUP, which was a between-subjects factor with two levels, SRC-ONLY and DORC-ONLY; and TRIAL TYPE, a within-subjects factor with two levels, TRAINED and UNTRAINED. Specifically, TRAINED stimuli were novel images with arrows eliciting subject relative clauses for the SRC-ONLY group and direct object relative clauses for the DORC-ONLY group, while UNTRAINED stimuli were also novel images, but with arrows eliciting direct object relative clauses for the SRC-ONLY group and subject relative clauses for the DORC-ONLY group.

Materials. Materials were the same as those in Experiment 1a except for the verbs provided during the test phase. To prevent participants from producing passive structures, verbs were presented in unambiguously active-voice forms (e.g., "was hugging").

Procedure. The procedure was identical to that of Experiment 1a except that during the test phase, if a participant produced the verb in a different form than that on the screen, experimenters asked participants to try again, using the verb as written. No feedback was given on the basis of the well-formedness of the response.

Results

Results are depicted in Figure 4 and summarized in Table 5. The final model converged with the full random effects structure after random effects correlations were removed. While DORC-ONLY participants still produced numerically fewer trained structures than SRC-ONLY participants, this main effect was no longer significant. The SRC-ONLY group produced significantly fewer untrained structures than trained ($\chi^2(1) = 21.287$, $p < .001$). Critically, the interaction representing the asymmetry in generalization was not significant as it had been in Experiment 1a ($\chi^2(1) = 0.373$, $p = .542$).

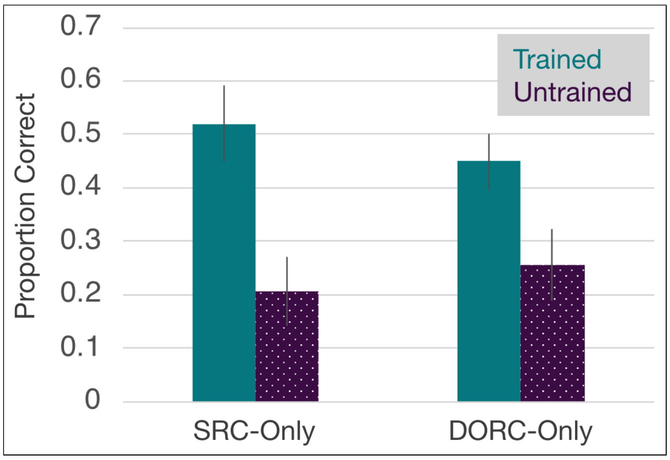


Figure 4. Proportion well-formed responses of the elicited structure as a function of GROUP and TRIAL TYPE in Experiment 1b.

To test whether generalization in this experiment reflected guessing behavior, we again compared participants who produced the TRAINED structure correctly more than 50% of the time to those who did not (Model 2; Figure 5 shows the relationship between learning of trained and untrained structures across participants). The model converged with the full random effects structure. A significant main effect of SUBSET confirmed that the 24 higher-performing participants generalized to the untrained structure (55% for the SRC-ONLY group and 63% for the DORC-ONLY group) more than the 24 low-performing participants (14% for SRC-ONLY and 12% for DORC-ONLY; model comparison confirmed that SUBSET significantly contributed to model fit: $\chi^2(1) = 6.255$, $p = .012$). This finding did not vary by group, and again indicates that generalization, at least among the high-performers, reflects above-chance performance.

Discussion

Experiment 1b removed a confound from Experiment 1a – namely, that participants in the SRC-ONLY group were able to produce either a trained structure or an untrained structure to correctly respond to DORC-TRIAL trials. We reasoned that if the asymmetry in generalization observed in Experiment 1a reflected this confound, it

Table 5

Experiment 1b statistical results.

	Model results			
	β	z	p	
Model 1: ONLY *groups* × TRIAL TYPE				
Intercept	0.050	0.114	.909	
GROUP: DORC-ONLY	−0.472	−0.778	.436	
TRIAL TYPE: UNTRAINED	−2.690	−4.812	< .001	***
Interaction	0.477	0.615	.538	
Model 2: ONLY *groups'* UNTRAINED *trials, split by performance on* TRAINED				
Intercept	−4.594	−3.552	< .001	***
GROUP: DORC-ONLY	−0.077	−0.052	.959	
SUBSET: >50%	3.040	2.038	.033	*
Interaction	1.520	0.867	.356	

would disappear in Experiment 1b where participants were prevented from producing passive SRCs in response to DORC-eliciting pictures. Indeed, while the SRC-ONLY group produced numerically fewer DORCs (20.6%) than the DORC-ONLY group produced SRCs (25.5%), this difference was no longer significant.

Experiments 1a and 1b therefore demonstrate that participants trained on only one type of relative clause are able to generalize to untrained types. This implicates the existence of a general representation of relative clauses, and thereby supports the Single Representation Hypothesis.

Experiment 2 served a number of purposes. First, we aimed to test the stability and abstractness of the general representation used to produce untrained structures. The experiment was also designed to increase statistical power. This followed from the observation that some features of the data differed from Experiment 1a to 1b in unexpected ways. For example, DORC-ONLY participants produced only 33.3% DORCs in Experiment 1a, but 44.9% in Experiment 1b. We therefore aimed to increase power in two ways. First, we increased the number of participants from 24 per group to 36. Second, following the Experiment 1a data indicating that training on syntactically diverse inputs might improve learning (i.e., the BOTH group numerically outperformed ONLY groups on TRAINED structures, despite half the training), we trained participants on two types of relative clauses and tested on a third. To do so, we used dative

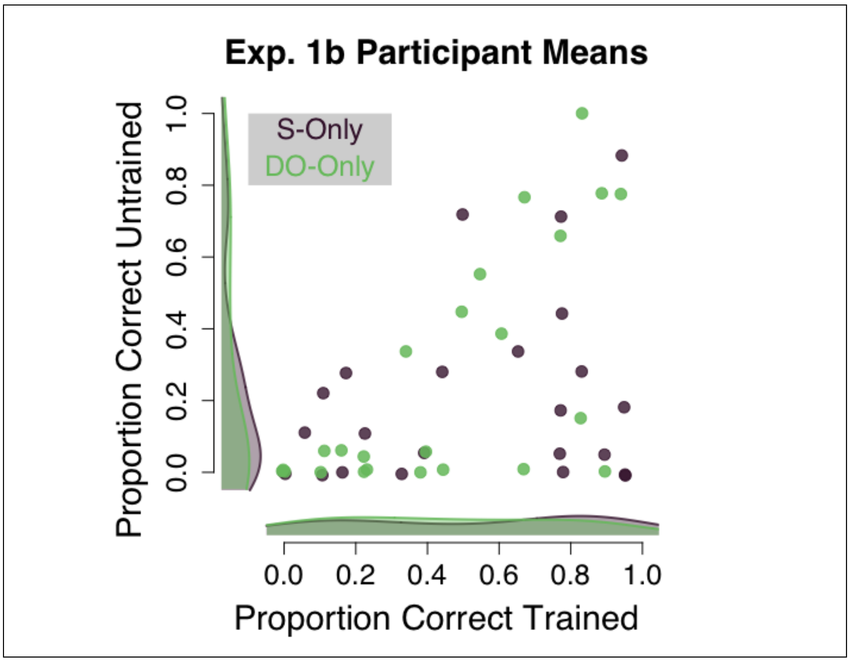


Figure 5. Experiment 1b: The relationship between learning trained structures and generalization to untrained structures.

constructions, from which three types of relative clauses can be formed: SRCs, DORCs, and IORCs.

Finally, the Experiment 2 grammar was designed to be more typologically typical than that of Experiments 1a and 1b. The clause-initial complementizer *that* was removed, as these are unattested among languages with prenominal relative clauses. The verb-final word order also makes the new grammar more typologically typical: roughly half of languages with verb-final dominant word order have prenominal relative clauses, while only five languages with verb-medial word order (as in English) have prenominal relative clauses (and all of these are dialects of Chinese or languages in close geographic proximity to Chinese-speaking populations; Comrie, 2008). If participants generalize in this grammar, it is more unlikely to reflect the use of knowledge of English.

Experiment 2

In Experiment 2, three groups of participants were trained on subsets of an artificial language that had prenominal relative clauses with verb-final word order. Participants received exposure to combinations of two types of relative clauses: SRCs and DORCs, DORCs and IORCs, or SRCs and IORCs. After training, participants were tested on their knowledge of all three types. Having increased statistical power and typological validity, we again ask whether participants will be able to produce structures that they have not been directly exposed to. If so, then it is further evidence for a general representation of relative clauses.

Method

Participants. Participants were continuously run until the target of 36 per group was reached. A total of 136 UC San Diego undergraduates participated for course credit. There were 28 exclusions: 5 for either being native speakers of a non-English language or for having learned Japanese (a language with finite prenominal relative clauses and verb-final word order), 7 due to experimenter error, and an additional 16 who experimenters reported were unable to learn the trained structures during the training phases. No exclusions were made on the basis of performance in the test phase.

Factors. Two factors were manipulated: TRIAL TYPE, which was a within-subjects factor with two levels: TRAINED and UNTRAINED; and GROUP, a between-subjects factor with three levels: SRC+DORC, SRC+IORC, and DORC+IORC.

Procedure. The procedure differed from that of Experiments 1a and 1b in two ways. First, to incrementally introduce novel grammatical properties to participants, we added a training phase to the beginning of the experiment in which participants learned to produce monoclausal sentences with the verb-final word order; 14 of these were monotransitive (i.e., including a subject and a direct object) and 14 were ditransitive (i.e., including a subject, a direct object, and an indirect object). This phase proceeded in the same way as the first training phase in previous experiments: experimenters described two pictures in a row, and then participants saw the two pictures in random order and recalled the descriptions, receiving help in the form of repeated full sentences as needed.

The second difference was in the test phase. Each trial consisted of three parts. First, an image of a ditransitive event appeared on the screen with no arrows. The participant was asked to produce a monoclausal sentence like the ones learned in the first training phase. Experimenters were instructed to help participants as necessary, ensuring that they had produced a monoclausal sentence with the correct meaning and with *Subject, Indirect Object, Direct Object, Verb* word order before proceeding.

Next, participants saw the same image, but with an arrow pointing to either the subject, direct object, or indirect object. Participants were instructed to produce an English translation of the relative clause sentence they would be asked to produce in the next part of the trial, for instance, "Those are the cookies that the grandmother baked for the children." Experimenters were again instructed to help as needed, and not to go on until the participant had produced a grammatical English relative clause describing the scene. By first asking participants to produce the monoclausal base sentence and then the English relative clause, we ensured that if participants did not produce a well-formed relative clause, it could not be attributed to a misunderstanding of the event or not knowing the base structure of the clause.

Finally, participants saw the same picture with the same arrow a second time and were asked to produce a sentence in their new grammar describing the person/animal/object the arrow pointed to. Experimenters gave no feedback during this part of the trial. The whole experiment lasted roughly two hours.

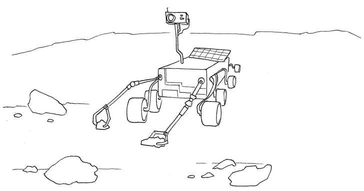
Materials. To reduce the length of the study, the number of items was reduced in training phases from 36 to 28. Participants were again tested on 36 new items in the test phase: 12 SRC-eliciting images, 12 DORC-eliciting images, and 12 IORC-eliciting images.

The artificial language in Experiment 2 used English words and had prenominal relative clauses and a verb-final word order. As in Experiments 1a and 1b, the internal structure of nominal elements (including DPs, NPs, and PPs) and all morphology (including verb tense and agreement) were identical to English. Training items contained several plural nouns in various syntactic positions and instances of verbs agreeing with singular and plural subjects to ensure that participants had cues to this effect.

Stimuli in the first training phase consisted of pictures of transitive and ditransitive events accompanied by simple monoclausal descriptions, as in Table 6. Word order for transitive events was always *Subject, Direct Object, Verb.* For ditransitive events, it was *Subject, Indirect Object, Direct Object, Verb.* Indirect objects were realized as prepositional objects with either the preposition *to* or *for*, as determined by the verb's preference in English (e.g., "give cookies *to*" but "bake cookies *for*").

Table 6

Experiment 2: sample stimulus items from Training Phase 1.

Monotransitive	Ditransitive
	
"[A robot] [moon rocks] collects."	"[The man] [for the woman] [a rose] bought."

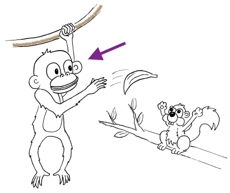
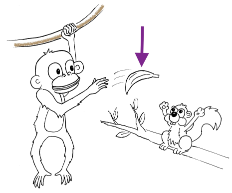
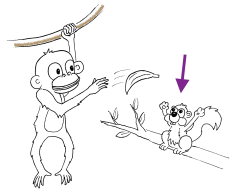
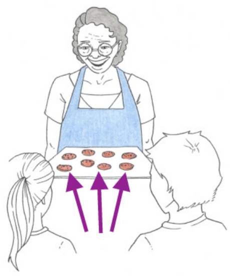
Stimuli in the second and third training phases consisted of pictures of ditransitive events with an arrow pointing to the subject, direct object, or indirect object. Pictures were paired with sentences with prenominal relative clauses, as in Table 7. Indirect object relative clauses were realized with a gap inside the prepositional phrase (so-called *adposition stranding*). In the test phase, stimuli consisted of images paired with verbs; sentences with prenominal relative clauses were the target elicitations.

Data coding and analysis. The most common errors are summarized in Table 8. To be coded as correct, a response had to be a well-formed prenominal relative clause of the elicited type (e.g., a SRC for an image with an arrow pointing to the subject) with the correct meaning.[11] Specific requirements included that the response contained a matrix clause with a subject, verb, and determiner that agreed in number with the head noun (e.g., "These$_{pl}$ are$_{pl}$ the$_{pl}$ [...] cookies$_{pl}$"); the internal structure of all nominal elements was correct in English (e.g., "the elderly couple"); the prenominal

[11] See Appendix C for an alternative coding scheme suggested by a reviewer and the results of analysis of those data.

Table 7

Experiment 2: sample relative clause stimulus items from Training Phases 2 and 3 and the Test Phase.

	SRC	DORC	IORC
Train			
	“That’s the [__ to the squirrel the banana threw] monkey.”	“That’s the [the monkey to the squirrel __ threw] banana.”	“That’s the [the monkey to __ the banana threw] squirrel.”
Test			
	bakes	bakes	bakes
Target	*That’s the [__ for the kids cookies bakes] grandma.*	*Those are [the grandma for the kids __ bakes] cookies.*	*Those are the [the grandma (for) __ cookies bakes] kids.*

relative clauses were finite and had *Subject, Indirect Object, Direct Object, Verb* word order (allowing for the gap position); and verbs agreed in number with subjects (e.g., “. . . the [the grandma$_{sg}$ for the kids __ bakes$_{sg}$] cookies”).

This latter requirement may have imposed a different level of difficulty on different training groups given the variable difficulty of maintaining noun-verb number agreement across different numbers and types of arguments. However, as can be seen in Table 8, the groups produced comparable numbers of this type of error. The DORC+IORC group may have been an exception, although their difficulty appeared in the trained structures, and not in untrained structures, indicating that this does not account for their lack of generalization.

The high number of missing prepositions in IORCs may reflect a reasonable

Table 8

Percentage of the most common errors in Experiment 2, by condition.

	SRC+DORC		SRC+IORC		DORC+IORC		
	Trained	Untr.	Tr.	Untr.	Tr.	Untr.	Overall
Constituent order error	17.13	23.15	12.50	40.24	3.36	40.97	18.90
Missing IORC preposition*	5.32	63.43	8.10	6.43	2.27	3.24	11.73
Repeated head determiner	9.38	14.35	7.26	11.19	5.44	5.56	8.36
Missing head determiner	7.52	8.56	6.19	6.19	5.90	3.47	6.39
Number agreement error	3.24	4.86	3.57	3.57	7.29	3.01	4.41
Wrong RC type	2.20	2.31	2.74	7.38	2.66	12.04	4.10

* While these were different from the input grammar, they are counted as correct in all Exp. 2 analyses.
Note 1. Errors were not mutually exclusive. Errors from responses that had three or more errors are not reported.
Note 2. Example correct response: "...the [to the squirrel the banana threw] monkey." Example constituent order error: "...the [the banana to the squirrel threw] monkey." Example missing preposition in IORC: "...the [the squirrel the banana threw] monkey." Example repeated head determiner: "...the [the squirrel the banana threw] the monkey." Example missing head determiner: "That's [to the squirrel the banana threw] monkey." Example number agreement error: "Here's a [the grandma for the kids baked] cookies."

strategy for relativizing a prepositional argument. Indeed, adposition stranding is exceedingly rare cross-linguistically. (Hungarian is, to our knowledge, the only language outside the Germanic family that allows this; Marácz, 1984.) A more common strategy is to simply delete the preposition, something which is done in a number of prenominal relative clause languages across families (e.g., Akhvakh, Evenki, Korean, Malayalam, Conchucos Quechua, etc.; Wu, 2011). We therefore coded IORCs as correct with either missing noun phrases or missing prepositional phrases.

Results

Numerical results are depicted in Figure 6 and broken down by structure type in Table 9. Statistical results are summarized in Table 10. The converging model contained random intercepts for participants and items and a random slope for TRIAL TYPE that varied within participants. There were no significant differences in how well the training groups learned the TRAINED structures. Untrained structures, however, were produced significantly less well than trained structures ($\chi^2(1) = 5.692$, $p = .017$). The SRC+DORC group produced more UNTRAINED structures than either of the other two groups (model comparison confirmed that the interaction term contributed to overall model fit: $\chi^2(1) = 56.349$, $p < .001$). To determine whether the SRC+IORC group also produced more UNTRAINED structures than the DORC+IORC group, we performed one additional pairwise comparison with a two-tailed test of equal proportions; this test confirmed the difference ($\chi^2(1) = 47.064$, Bonferroni-corrected-$p < .001$).

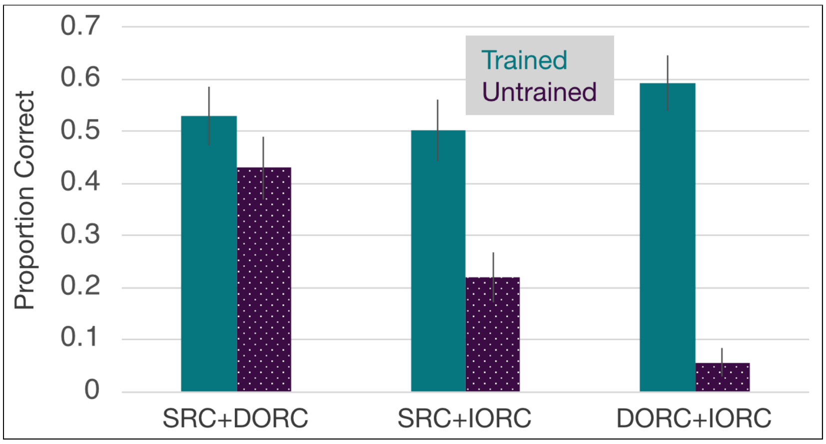


Figure 6. Proportion of well-formed productions of the elicited structure as a function of GROUP and TRIAL TYPE in Experiment 2.

The relationship between learning and generalization across participants is shown in Figure 7. To determine whether production of untrained structures reflected educated guessing, we again compared participants who produced more than 50% of trained structures to those who produced fewer. The converging model contained random intercepts for participants and items and a random slope for group that was allowed to vary within participants. A significant effect of SUBSET confirmed that the 64 higher-performing participants generalized more (SRC+DORC: 75%, SRC+IORC: 67%, DORC+IORC: 53%) than the 43 lower-performing participants (SRC+DORC: 18%, SRC+IORC: 16%, DORC+IORC: 6%; model comparison: $\chi^2(1) = 22.732$, $p < .001$). There was no statistical evidence that this effect varied by group. This again indicates that generalization does not reflect guessing, but syntax-driven production.

Table 9

Experiment 2: proportion correct responses by group and structure type.

	TRIAL TYPE		
GROUP	SRC	DORC	IORC
SRC+DORC	.501	.560	.431
SRC+IORC	.482	.220	.524
DORC+IORC	.056	.588	.598

Discussion

Experiment 2 differed from Experiments 1a and 1b in that it had greater statistical power and employed a grammar involving a different basic word order. We replicated the finding from Experiments 1a and 1b that participants produced structures they had not been directly exposed to. This supports the idea of a single, general representation of relative clauses, and suggests that this representation is stable and abstract enough to accommodate a variety of grammars.

The particular pattern of generalization in Experiment 2, however, was unexpected. Participants in the SRC+DORC group generalized to IORCs and participants in the SRC+IORC group generalized to DORCs, but very few participants in the DORC+IORC generalized to SRCs. To determine whether this might reflect some specific feature of the grammar in Experiment 2, Experiment 3 was designed to replicate Experiment 2 using a different grammar. To simplify the design, we did not include an SRC+IORC group.

Experiment 3

Experiment 3 aimed to replicate the unexpected pattern of generalization displayed by the SRC+DORC and DORC+IORC groups in Experiment 2. We again modified the grammar so as to ensure that our findings were not somehow an artifact of properties of the particular artificial language. The new grammar was a modified version of the Experiment 2 grammar; it had bare nouns (i.e., with no determiners) and case marking (the use of suffixes to specify whether a noun is a subject, direct object, indirect object, oblique object, etc., as in, e.g., Korean). In a continued effort to adhere

Table 10

Experiment 2 results.

	Model results			
	β	z	p	
Model 1: GROUP × TRIAL TYPE				
Intercept	0.088	0.229	.819	
GROUP: SRC+IORC	−0.249	−0.460	.646	
GROUP: DORC+IORC	0.214	0.460	.689	
TRIAL TYPE: UNTRAINED	−0.810	−2.406	.016	*
Interaction: SRC+IORC, UNTRAINED	−1.694	−3.334	< .001	***
Interaction: DORC+IORC, UNTRAINED	−4.606	−7.418	< .001	***
Model 2: UNTRAINED *trials, split by performance on* TRAINED				
Intercept	−2.853	−4.775	< .001	***
GROUP: SRC+IORC	−1.752	−1.854	.064	.
GROUP: DORC+IORC	−21.12	−0.001	.999	
SUBSET: >50%	3.852	5.065	< .001	***
Interaction: SRC+IORC, UNTRAINED	0.033	0.028	.978	
Interaction: DORC+IORC, UNTRAINED	15.83	0.001	.999	

to typological typicality and to reduce reliance on knowledge of English, we removed prepositions from the grammar (which are more commonly realized as postpositions and/or case-marking in verb-final languages).

Method

Participants. Participants were continuously run until a pre-determined stop date or the target of 36 per group was reached. Because the stop date was reached prior to reaching the target number, a total of 45 participants were run. Data from 3 participants were excluded: 1 for early childhood exposure to Chinese (which has prenominal relative clauses), 1 for attempting to explicitly describe the grammar aloud throughout the training phases, and 1 who was unable to learn the trained structures. No exclusions were made on the basis of performance in the test phase. The final data set included 21 participants in each group.

Factors. Two factors were manipulated: TRIAL TYPE, which was a within-subjects factor with two levels, TRAINED and UNTRAINED; and training GROUP, a between-subjects factor with two levels: SRC+DORC and DORC+IORC.

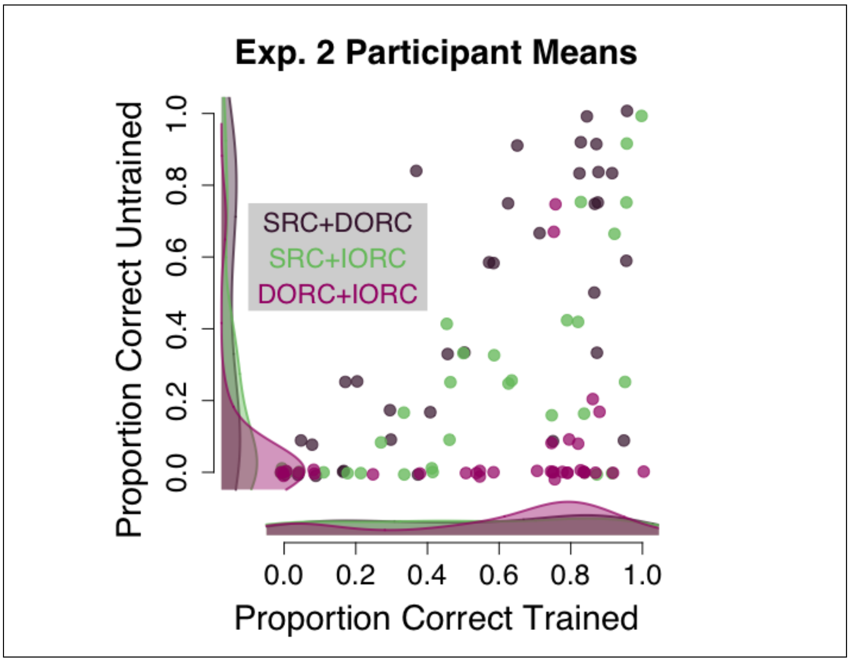


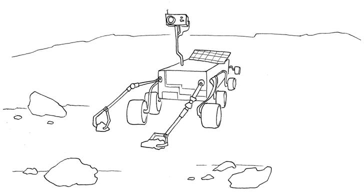
Figure 7. Experiment 2: The relationship between learning trained structures and generalization to untrained structures. The fact that participants in all three groups performed comparably on TRAINED trials can be seen in the relatively even spread of participants in each group along the horizontal axis (see density plots along bottom axis). The very weak correlation for the DORC+IORC group reflects the fact that this group generalized much less than either other two groups. This can also be seen in the density plot along the vertical axis: participants are clustered at the bottom.

Materials. In response to experimenters reporting that participants appeared to fatigue toward the end of the training phases, we again shortened the experiment by reducing the number of training items from 28 to 20: 10 of each of the trained structures. Participants were again tested on 36 new items, 12 of which elicited SRCs, 12 DORCs, and 12 IORCs.

The grammar of the artificial language in Experiment 3 was a modified version of the Experiment 2 grammar. It had no determiners, no prepositions, and included nominative, accusative, and dative case marking suffixes (although the head noun was not marked for case, consistent with the way case-marking languages like Korean and Japanese treat the object of the verb *be*). These changes removed many of the remaining components of English word order from the relative clause. The grammar was more typologically typical in that it included case marking and excluded prepositions, both of which are common features of verb-final languages. Sample items appear in Tables 6 and 7.

Table 11

Experiment 3: Sample stimulus items from Training Phase 1.

Monotransitive	Ditransitive
	
"Robot-uh rocks-en collects."	"Man-uh woman-ik rose-en buys."

Procedure. The procedure was identical to that of Experiment 2. The whole experiment took roughly 90 minutes.

Data coding. Responses were coded as correct if they contained a prenominal relative clause that conformed to the input grammar.[12] Specific criteria included that nouns inside the relative clause had the correct case markers and appeared in the correct order: *Subject, Indirect Object, Direct Object, Verb* (modulo the missing noun). The head noun was not allowed to have case marking. The most frequent errors are reported in Table 13.

Results

Data are shown in Figure 8 and broken down by structure in Table 14. Models are reported in Table 15. The main analysis, Model 1, converged with random intercepts for items and participants, no random effects correlations, a random slope for TRIAL TYPE that varied within participants, and random slopes for TRIAL TYPE and the TRIAL TYPE×GROUP interaction that varied within items. Neither the main effect of GROUP nor the main effect of TRIAL TYPE were significant, but their interaction was

[12] See Appendices B and C for different coding schemes suggested by reviewers and the results of statistical analyses using those data.

Table 12

Experiment 3: Sample relative clause stimulus items from Training Phases 2 and 3 and the Test Phase.

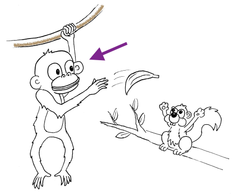
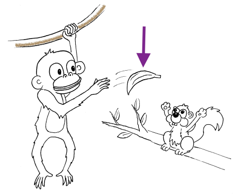
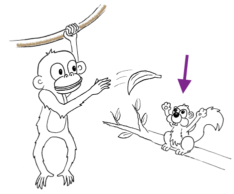
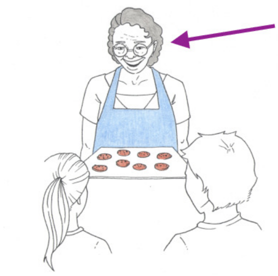
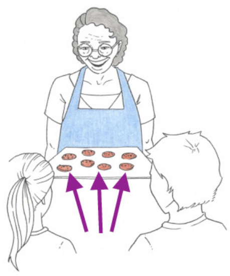
	SRC	DORC	IORC
Train			
	"Here's this [__ squirrel-ik banana-en throws] monkey."	"Here's this [monkey-uh squirrel-ik __ throws] banana."	"Here's this [monkey-uh __ banana-en throws] squirrel."
Test			
	bakes	bakes	bakes
Target	*Here's this [__ kids-ik cookies-en bakes] grandma.*	*Here are these [grandma-uh kids-ik __ bakes] cookies.*	*Here are these [grandma-uh __ cookies-en bakes] kids.*

Table 13

Percentage of the most common errors in Experiment 3, by condition.

	SRC+DORC		DORC+IORC		
	Trained	Untrained	Trained	Untrained	Overall
Constituent order error	36.51	29.71	23.41	27.38	29.50
Wrong case marker	7.54	17.46	3.17	15.87	9.13
Missing case marker	12.70	16.67	2.58	5.95	8.86
Case marker on head noun	6.94	10.32	3.17	10.32	6.81
Number agreement error	2.18	6.35	4.17	1.19	3.37

Note 1. Errors were not mutually exclusive. Errors from responses that had three or more errors are not reported in this table.
Note 2. Example correct SRC response: "…this [squirrel-ik banana-en throws] monkey." Example constituent order error: "…this [banana-en squirrel-ik throws] monkey." Example wrong case marker: "…this [squirrel-uh banana-en throws] monkey." Example missing case marker: "…this [squirrel-ik banana throws] monkey." Example case marker on head noun: "…this [squirrel-ik banana-en throws] monkey-uh." Example number agreement error: "Here's this [grandma-uh kids-ik bakes] cookies."

($\chi^2(1) = 7.173$, $p = .007$), reflecting the fact that the SRC+DORC group generalized more than the DORC+IORC group.

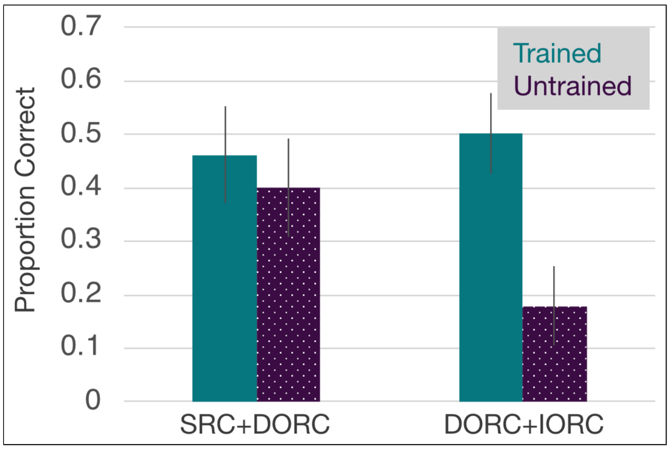


Figure 8. Proportion of well-formed productions of the elicited type as a function of GROUP and TRIAL TYPE in Experiment 3.

The relationship between learning and generalization across participants is shown in Figure 9. We compared the amount of generalization among the 20 learners who produced more than 50% correct trained structures to the 22 who did not. The model, Model 2 in Table 15, converged with a full random effects structure but no random effects correlations. Higher-performing participants generalized more (81% in the SRC+DORC group and 64% in the DORC+IORC group) than lower-performing participants (11% in the SRC+DORC group and 17% in the DORC+IORC group; $\chi^2(1) = 12.78$, $p < .001$), again indicating that generalization does not reflect guessing.

Discussion

Experiment 3 replicated the findings of previous experiments in that participants produced the UNTRAINED structures. It furthermore replicated the specific pattern of generalization observed in Experiment 2, whereby SRC+DORC participants generalized to IORCs more than DORC+IORC participants generalized to SRCs. Across groups, generalization was higher for participants that learned the trained structure better, indicating that generalization, at least among the highest performing participants, does not reflect guessing.

Table 14

Experiment 3: proportion correct responses by group and structure type.

	TRIAL TYPE		
GROUP	SRC	DORC	IORC
SRC+DORC	.468	.456	.401
DORC+IORC	.179	.421	.583

General Discussion

The present goal has been to understand the representation of long-distance syntactic dependencies. Specifically, we tested two competing hypotheses about the underlying representation of relative clauses. The Multiple Representations Hypothesis posits that there are many specific, surface-oriented representations, whereas the Single Representation Hypothesis suggests that there is one representation that is general enough to account for all the variety in surface word orders. To distinguish between these possibilities, we trained participants on only some types of relative clauses and tested whether they also learned other types, despite not having had direct exposure to them. If not, it would be an indicator that each type of relative clause is represented independently. But if so, it would support a model of different types of relative clauses as having the same underlying representation: something like a general principle according to which a head noun is not repeated inside the relative clause.

In four experiments, participants were implicitly trained on a new grammar for relative clauses. Experiments 1a and 1b provided the first evidence that learners of one type of relative clause also learn other types. In Experiment 1a, one group received an explicit grammar lesson rather than implicit training. At test, this group performed worse than the implicit groups, indicating that the implicit groups' productions were not guided by explicit representations. However, the conclusion that participants learned untrained structures is called into question by the fact that many passive relative clauses were produced, perhaps indicating that participants used knowledge of English syntax to perform the task.

Experiments 2 and 3 removed this confound by teaching grammars that were even more different from English so as to reduce the chances that generalization might rely on

Table 15

Experiment 3 results.

	Model results			
	β	z	p	
Model 1: GROUP × TRIAL TYPE				
Intercept	−0.367	−0.513	.609	
GROUP: DORC+IORC	0.280	0.281	.779	
TRIAL TYPE: UNTRAINED	−1.238	−1.435	.151	
Interaction	−3.295	−2.624	.009	**
Model 2: UNTRAINED *trials, split by performance on* TRAINED				
Intercept	−6.681	−3.268	.001	**
GROUP: DORC+IORC	−1.300	−0.598	.550	
SUBSET: >50%	9.421	2.963	.003	**
Interaction	−5.979	−1.309	.191	

English knowledge. Both experiments showed evidence for generalization. This finding implies that when the syntax of a relative clause is learned, it is not a representation of a specific structure, but a more abstract representation of relative clauses in general.

We have suggested that this general representation may take the form of a principle like, "Leave out the head noun inside the relative clause." An important observation is that this principle also holds true for English. Despite the fact that the novel grammars were different from English (and increasingly so across experiments), participants in our study may have simply learned a new basic word order (e.g., in Experiments 2 and 3: *Subject, Direct Object, Indirect Object, Verb*), but used pre-existing knowledge of an English general relative clause representation to generate untrained types. As pointed out in the introduction, it does not matter whether the general representation was learned in the training phase or borrowed from English. Either supports the Single Representation Hypothesis because participants used a single, general representation of relative clauses.

It is also worth noting that, in each experiment, not all participants generalized to the untrained structure. This is not unexpected, given the noisiness of behavioral data in general and language-production data in particular. For example, there may be wide variation in how much attention individual participants dedicate to the experimental tasks, and participants who attend less may acquire untrained structures less than

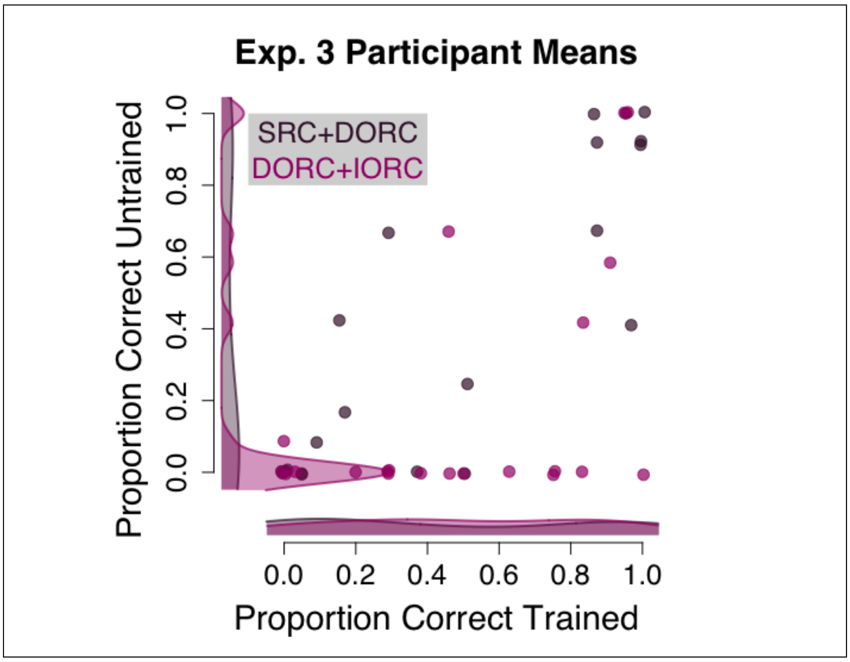


Figure 9. Experiment 3: The relationship between learning trained structures and generalization to untrained structures.

participants who attend more (indeed, this is consistent with our analyses showing that participants who learn trained structures better also learn untrained structures better). Given evidence across these experiments that participants provided with exposure to trained structures tended to produce more untrained structures than expected by guessing, the conclusions drawn about the operation of a single representation remains warranted. (Variation in propensity to produce untrained structures more systematically deserve further comment – see below, where we discuss why the DORC+IORC groups in Experiments 2 and 3 may not have produced many SRCs.)

The results presented here constitute the first experimental evidence that we are aware of for a general representation of long-distance dependencies like relative clauses. This finding indicates that the taxonomy of syntactic representations includes at least two levels of abstraction: one over words and phrases, and another over this level. This corroborates the assumption of theoretical approaches, which have long aimed to unify this class of structures (Chomsky, 1981; Culicover & Jackendoff, 2005; Pollard & Sag, 1994).

Modern psychological approaches also assume such a level of abstraction, however it is not clear that these are intended to include specific representations like that in (3), reproduced here:

(7) Head Noun [*that* Subject Verb (Direct Object) (...)]**

**DO NOT REPEAT THE HEAD NOUN

That is, *neighborhoods* of representations are stipulated in order to formally capture the similarity among representations like SRCs, DORCs, and other relative clauses (Goldberg, 1995; Y. Hsiao & MacDonald, 2013; Juliano & Tanenhaus, 1993; Kim, Srinivas, & Trueswell, 2002). However, similarity alone is not enough to account for how a speaker might generate a IORC having never encountered one. This finding implicates a representation that is specific enough to guide production. Such a representation must be specific enough to encode that relative clauses consist of ordinary embedded clauses in which the head noun is omitted. This degree of specificity is not inconsistent with extant psychological models, but neither is it clearly assumed.

Lower performance on untrained structures

If knowledge of one relative clause is knowledge of all relative clauses, then we might have expected performance to be equally good on untrained structures and trained structures. However, across experiments, participants consistently performed better on trained structures than untrained ones. This indicates that these structures were still in some sense novel to participants.

There are a number of possibilities for why this might be. One is that it simply reflects a lack of practice in producing the untrained structure, but not an inability to do so. For example, participants trained on SRCs receive a good deal of practice relativizing agents. But relativizing patients, as in many of our DORC stimuli, may also require practice, which SRC-ONLY participants received far less of. Thus, even if participants have a general representation for relative clauses, they may need to practice mapping from particular types of semantic representations to the syntactic representation.

Another intriguing possibility – one that we consider likely – is that relative clauses may be simultaneously represented in multiple ways. That is, it is possible that learners acquire both a general representation, as we have argued on the basis of generalization, and a specific representation that directly reflects the particular surface word order of each type.

It has long been noted that general and specific representations are not mutually exclusive (Goldberg, 1995). Idioms, for instance, must have both compositional and noncompositional representations. Rachel can "pull Dan's leg" – or joke around with Dan – and Dan can "pull Rachel's leg" – Dan can joke around with Rachel. If these had no compositional structure, then each would have to be learned as a separate idiom to be correctly interpreted. But if they had no noncompositional representation, then they would be about two people tugging on limbs. It is possible that the difficulty associated with producing untrained structures may reflect the difficulty associated with deriving a specific representation from the general representation. (This may not in fact be different from the lack-of-practice possibility mentioned above.)

A final, related possibility is that syntactic structures can be generated online via analogy (Gentner, 1983). If relative clauses exist in neighborhoods defined by similarity, consistent with modern psychological approaches (Y. Hsiao & MacDonald, 2013; Juliano & Tanenhaus, 1993; Kim et al., 2002; Wells, Christiansen, Race, Acheson, & MacDonald, 2009), it stands to reason that speakers may typically rely on specific representations, but may use neighboring structures to analogize to new ones as needed.

Crucially, however, an analogy-based account is consistent with the Single Representation Hypothesis. That is, one cannot analogize from one thing to another without an abstract representation of the underlying similarity between the two. That is, generating a direct object relative clause from a subject relative clause via analogy requires you to know that relative clauses are subordinate clauses with the head noun left out. This knowledge is exactly the type of Single Representation our data point to.

Whatever the reason for the lower performance on untrained structures, there is still a need for a general representation of relative clauses to account for the fact that some participants reliably produced untrained structures in each of the four experiments. The idea of such a representation is not new. Indeed, Chomsky (1959) cited the need for general representations of long-distance dependencies in his famous challenge to behaviorism, and several theoretical frameworks attempting to formalize this idea have since been developed. For example, *derivational* models account for long-distance dependencies with operations that "move" or "delete" noun phrases (e.g., Chomsky, 1981, 1992), while more surface-oriented models include lexical rules which introduce "slash-categories," sharing features with disparate parts of the representation (e.g., Culicover & Jackendoff, 2005; Pollard & Sag, 1994; Sag, Wasow, Bender, & Sag, 1999).

To the best of our knowledge, all prior evidence for a general representation of long-distance dependencies has been inferred from linguistic observations such as the similarities between the various types of English relative clauses. Such data have the benefit of being derived from the behavior of native speakers, who learned languages in ways far more naturalistic than participants in our experiments. But it is also impossible to disentangle various influences on the trajectory of acquisition of such structures in a speaker's native language. As noted earlier, even very young children have probably been exposed to all types of relative clauses in their language. As such, it cannot be determined whether productive use of a given structure reflects generalization on the one hand, or reliance on direct experience on the other (although see Montag & MacDonald, 2015 for compelling evidence favoring direct experience).

The present study supplements the linguistic data with experimental data. The experiments having been carefully controlled, confounds such as the potential for exposure to other types of relative clauses are minimized and the observed generalization to untrained types constitutes strong evidence for the existence of a general underlying representation.

Asymmetries in generalization

Another interesting feature of our data was that in two experiments, participants trained on DORCs and IORCs did not generalize to SRCs as much as participants trained on SRCs and DORCs generalized to IORCs. This pattern is consistent with the order in which children acquire various relative clauses in their native languages. In particular, Yip and Matthews (2007) documented the acquisition of relative clauses in Cantonese-English bilingual children. They found that the first relative clauses children acquired in both languages were DORCs. (Curiously, these DORCs were prenominal in English as well as in Cantonese.)

From a cross-linguistic perspective, however, this order of acquisition is somewhat surprising. This is because of the existence of a typological contingency hierarchy, the *Noun Phrase Accessibility Hierarchy*, according to which the existence of DORCs and IORCs in a language implicates the existence of SRCs, but the existence of SRCs and DORCs does not necessarily implicate the existence of IORCs (Keenan & Comrie, 1977). That is, there are no languages that have DORCs and IORCs but do not have SRCs, but there are languages (e.g., Hebrew and Standard Arabic) which have SRCs and DORCs, but not IORCs. This is the opposite of the pattern that we observe.

The Noun Phrase Accessibility Hierarchy has long been thought to reflect processing difficulty (Keenan & Comrie, 1977; Keenan & Hawkins, 1987). That is, extensive research demonstrates that SRCs are easier to process than DORCs and IORCs (Clemens et al., 2015; Cook, 1975; Diessel & Tomasello, 2005; Hatch, 1971; Keenan & Hawkins, 1987; Kwon, Lee, Gordon, Kluender, & Polinsky, 2010; Wagers, Borja, & Chung, 2018), although most research has solely focused on differences between SRCs and DORCs. For instance, studies dating back decades have consistently found evidence that, relative to SRCs, DORCs are more difficult to process. Using a continuous lexical decision task, Ford (1983) showed that reading DORCs required significantly more time per word than reading SRCs, and Holmes and O'Regan (1981) used an eyetracking paradigm to demonstrate that reading DORCs leads to more regressions (looks back to previous words) than reading SRCs. Similarly, elicited production tasks commonly find a preference for SRCs over DORCs (modulo other relevant processing factors like relative animacy of nouns), (e.g., McDaniel, McKee, & Bernstein, 1998). Corpus work has confirmed this; in a corpus of American English with over 22 million words, Reali and Christiansen (2007) found that over 65% of relative clauses were SRCs (although see Duffield & Michaelis, 2011, for an alternative explanation of the frequency data).

These processing findings suggest that the Noun Phrase Accessibility Hierarchy may in fact be a sort of grammatical fossil, the result of impact of minor differences in difficulty associated with the various types of relative clauses played out over the course of language evolution. This would be consistent with a finding from the statistical learning literature that suggests that easier patterns come along for the ride when learners are trained on harder patterns (Thompson & Newport, 2007).

However, this cannot explain the pattern we observe in Experiments 2 and 3. Indeed, it predicts the opposite. One possible explanation is that difficulty may have different effects depending on the particular task. For instance, diachronically, it may lead to attrition of more difficult structures. In synchronic language learning, on the other hand, the higher difficulty of DORCs and IORCs may lead to a stronger focus on acquiring specific representations for these structures, ultimately distracting from the acquisition of a general representation of relative clauses (if in fact these two types of representation coexist). If this is the case, then where the trained structures are especially difficult to learn, the general representation is less well acquired and thus used less effectively. This would predict poor generalization, as was observed in Experiments 2 and 3.

Another possibility is that the asymmetrical patterns of processing difficulty are different in prenominal relative clause languages. Indeed, this is the prediction of Dependency Locality Theory (Gibson, 1998), according to which the difficulty of relative clause processing is a function of the distance between the head noun and the gap. Most previous research on processing asymmetries between types of relative clauses has been done on postnominal relative clause languages like English. In these languages, subject gaps are closer to the head noun than object gaps, so the theory (correctly) predicts that SRCs are easier to process than DORCs. However, in prenominal relative clause languages, provided that subjects precede objects in the basic word order (which is almost always the case), subject gaps are *farther* from the head noun than object gaps, predicting that they should be harder to process.

This prediction has been tested in a number of prenominal relative clause languages, and the evidence remains mixed. For instance, Kwon, Polinsky, and Kluender (2006) found faster reading for Korean SRCs relative to DORCs, and in an ERP study, Kwon, Kluender, Kutas, and Polinsky (2013) found evidence that SRCs are easier to process than DORCs in Chinese, Korean, and Japanese (all prenominal relative clause languages). Similarly, in a sentence-character matching study, Hu, Gavarró, Vernice, and Guasti (2016) found that Mandarin-speaking children comprehended SRCs more accurately than DORCs. However, F. Hsiao and Gibson (2003) demonstrate faster reading times in a self-paced reading paradigm for Chinese DORCs relative to SRCs. Gibson and Wu (2013) replicate this result, and show how it is possible to obtain the opposite result due to temporary ambiguities in stimuli that are not properly controlled, casting doubt on some findings showing a processing advantage for SRCs.

If prenominal DORCs are in fact easier to process than their SRC counterparts, it may go some distance to explaining the asymmetrical patterns of generalization we observed in Experiments 2 and 3. Specifically, we might expect that the success of generalizing to a new structure correlates with that structure's associated degree of processing difficulty.

However, two caveats are worth noting. First, if the Noun Phrase Accessibility Hierarchy is a result of the kinds of processing asymmetries we discuss here, then one would expect the hierarchy to reverse in languages with prenominal relativization. That is, if DORCs are easier to process than SRCs in a given set of languages, then one might expect that some of those languages would allow DORCs but not SRCs, and not the reverse. To our knowledge, such languages are unattested. Second, while our data are consistent with a different pattern of difficulty among relative clause types than that observed in English, they are not consistent with Dependency Locality Theory. This is because word order in Experiments 2 and 3 was *Subject, Indirect Object, Direct Object, Verb*. If distance between the head noun and the gap were the sole predictor of processing difficulty, then one would expect that DORCs should be learned the best in our study, followed by IORCs, and finally SRCs. This is not the pattern observed in Experiment 2, where IORCs were learned best of all, followed by DORCs, then SRCs (see Figure 6).[13]

Conclusion

We began by pointing out that there are two ways in which the syntax of long-distance dependencies might be represented: On the one hand, they may be represented as a family of independent representations with similar properties, although this would be a relatively uneconomical means of representation. On the other hand, they may have a single general representation. While this latter possibility may seem like an unnecessarily baroque type of representation to posit, it is in fact the way most theoretical accounts of syntax model long-distance dependencies.

Here, we present four experiments which carefully control exposure such that learners have direct experience with some types of relative clauses, but not with others. In spite of this limited exposure, they are able to produce structures that they have never been exposed to, indicating that they have access to a general representation. This finding indicates that syntactic knowledge has at least two levels of abstract representations above the level of words and phrases, corroborating extant theoretical and psychological models of syntax.

Data Availability Statement

The data that support the findings of this study are openly available on OSF at `https://osf.io/4VDRT/` under DOI 10.17605.

[13] Another possibility for explaining the generalization asymmetry was pointed out by an anonymous reviewer: This may be related to an *inverse preference effect* in the structural priming literature, whereby exposure to dispreferred structural alternates (e.g., a passive) can give rise to stronger priming than exposure to the preferred alternate (e.g., an active). It is possible that the asymmetrical generalization we observe in this study is due to the same mechanism.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*(2-3), 150–177.

Birner, B. J., & Ward, G. L. (1998). *Information status and noncanonical word order in english* (Vol. 40). John Benjamins Publishing.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about knowing*, *185*.

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning. *Linguistic perspectives on second language acquisition*, *4*, 1–68.

Bley-Vroman, R., Felix, S. W., & Loup, G. L. (1988). The accessibility of universal grammar in adult language learning. *Interlanguage studies bulletin (Utrecht)*, *4*(1), 1–32.

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review*, *3*(4), 395–421.

Bolinger, D. (1977). Meaning and form.

Bresnan, J., Asudeh, A., Toivonen, I., & Wechsler, S. (2015). *Lexical-functional syntax*. John Wiley & Sons.

Chomsky, N. (1959). *Chomsky, n. 1959. a review of bf skinner's verbal behavior. language, 35 (1), 26–58.* JSTOR.

Chomsky, N. (1981). *Lectures on government and binding* (No. 9). Walter de Gruyter.

Chomsky, N. (1992). *The minimalist program*. MIT press.

Christiansen, M. H. (2001). *Using artificial language learning to study language evolution: Exploring the emergence of word order universals*. na.

Christiansen, M. H., & Chater, N. (2015). The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology*, *6*, 1182.

Clemens, L. E., Coon, J., Pedro, P. M., Morgan, A. M., Polinsky, M., Tandet, G., & Wagers, M. (2015). Ergativity and the complexity of extraction: A view from mayan. *Natural Language & Linguistic Theory*, *33*(2), 417–467.

Comrie, B. (2008). Prenominal relative clauses in verb-object languages. *Language and Linguistics*, *9*(4), 723–733.

Cook, V. J. (1975). Strategies in the comprehension of relative clauses. *Language and Speech*, *18*(3), 204–212.

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, *6*(5), 310–329.

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

Culbertson, J., & Newport, E. L. (2017). Innovation of word order harmony across development. *Open Mind*, *1*(2), 91–100.

Culbertson, J., & Smolensky, P. (2012). A bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive science*, *36*(8), 1468–1498.

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

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax.* Oxford University Press on Demand.

Davis, H. (2010). A unified analysis of relative clauses in st'át'imcets. *Northwest Journal of Linguistics*, *4*(1), 1–43.

Diessel, H. (2004). *The acquisition of complex sentences* (Vol. 105). Cambridge University Press.

Diessel, H., Dąbrowska, E., & Divjak, D. (2015). Usage-based construction grammar.

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses.

Language, 882–906.

Dryer, M. S. (2013). Order of relative clause and noun. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/90`

Duffield, C. J., & Michaelis, L. A. (2011). Why subject relatives prevail: Constraints versus constructional licensing. *Language and Cognition*, *3*(2), 171–208.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79.

Eilish, B. (2019). *Bad guy.* Interscope Records.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, *32*(5), 429–448.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2016). Miniature artificial language learning as a complement to typological data. *The usage-based study of language learning and multilingualism*, 211–232.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 501–538.

Fitz, H., Chang, F., & Christiansen, M. H. (2011). A connectionist account of the acquisition and processing of relative clauses. *The acquisition of relative clauses*, *8*, 39–60.

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 203–218.

Fox, B. A. (1987). The noun phrase accessibility hierarchy reinterpreted: Subject primacy or the absolutive hypothesis? *Language*, 856–870.

Gass, S. (1979). Language transfer and universal grammatical relations. *Language learning*, *29*(2), 327–344.

Gass, S., & Ard, J. (1980). L2 data: Their relevance for language universals. *TESOL Quarterly*, 443–452.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*(1), 1–23.

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive psychology*, *65*(2), 141–176.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155–170.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gibson, E., & Wu, H.-H. I. (2013). Processing chinese relative clauses in context. *Language and Cognitive Processes*, *28*(1-2), 125–155.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2002). Surface generalizations: An alternative to alternations. *Cognitive Linguistics*, *13*(4), 327–356.

Goldberg, A. E. (2013). Constructionist approaches. *The Oxford handbook of construction grammar*, *1*.

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective onalternations'. *International journal of corpus linguistics*, *9*(1), 97–129.

Hatch, E. (1971). The young child's comprehension of relative clauses.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, *45*(3), 718–730.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*(4), 417–430.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in chinese. *Cognition*, *90*(1), 3–27.

Hsiao, Y., & MacDonald, M. C. (2013). Experience and generalization in a connectionist model of mandarin chinese relative clause processing. *Frontiers in psychology*, *4*, 767.

Hu, S., Gavarró, A., Vernice, M., & Guasti, M. T. (2016). The acquisition of chinese relative clauses: contrasting two theoretical approaches. *Journal of child language*, *43*(1), 1.

Humphreys, G. F., Mirković, J., & Gennari, S. P. (2016). Similarity-based competition in relative clause production and comprehension. *Journal of Memory and Language*, *89*, 200–221.

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic bulletin & review*, *23*(4), 1214–1220.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, *132*(2), EL74–EL80.

Juliano, C., & Tanenhaus, M. K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the 15th annual conference of the cognitive science society* (pp. 593–598).

Kako, E. (1999). Elements of syntax in the systems of three language-trained animals. *Animal Learning & Behavior*, *27*(1), 1–14.

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

Kang, S. H., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic* *bulletin & review*, *20*(6), 1259–1265.

Kaplan, R. M., & Zaenen, A. (1989). Long-distance dependencies, constituent structure, and functional uncertainty. *Alternative conceptions of phrase structure*, *17*, 42.

Kaschak, M. P. (2006). What this construction needs is generalized. *Memory & cognition*, *34*(2), 368–379.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic inquiry*, *8*(1), 63–99.

Keenan, E. L., & Hawkins, S. (1987). The psychological validity of the accessibility hierarchy. *Universal grammar*, *15*, 60–85.

Kim, A. E., Srinivas, B., & Trueswell, J. C. (2002). A computational model of the grammatical aspects of word recognition as supertagging. *The lexical basis of sentence processing: Formal, computational and experimental issues*, *4*, 109.

Kwon, N., Kluender, R., Kutas, M., & Polinsky, M. (2013). Subject/object processing asymmetries in korean relative clauses: Evidence from erp data. *Language*, *89*(3), 537.

Kwon, N., Lee, Y., Gordon, P. C., Kluender, R., & Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of prenominal relative clauses in korean. *Language*, 546–582.

Kwon, N., Polinsky, M., & Kluender, R. (2006). Subject preference in korean. In *Proceedings of the 25th west coast conference on formal linguistics* (pp. 1–14).

Langacker, R. W. (2009). Constructions and constructional meaning. *New directions in cognitive linguistics*, 225–267.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of child language*, *30*(2), 333.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology*, *4*, 226.

Marácz, L. K. (1984). Postposition stranding in hungarian. *GAGL: Groninger Arbeiten zur germanistischen Linguistik*(24), 127–161.

Martin, A., Abels, K., Adger, D., & Culbertson, J. (2019). Do learners' word order preferences reflect hierarchical language structure? In *Cogsci* (pp. 2303–2309).

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The cappuccino model. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

McDaniel, D., McKee, C., & Bernstein, J. B. (1998). How children's relatives solve a problem for minimalism. *Language*, 308–334.

McElree, B., & Bever, T. G. (1989). The psychological reality of linguistically defined gaps. *Journal of Psycholinguistic Research*, *18*(1), 21–35.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447.

Ochs, E., & Schieffelin, B. B. (1994). Language acquisition and socialization: Three developmental stories and their implications. In B. G. Blount (Ed.), *Language, culture, and society* (p. 470-512). Prospect Heights, IL, USA: Waveland Press, Inc.

Pepperberg, I. M. (1981). Functional vocalizations by an african grey parrot (psittacus erithacus). *Zeitschrift für Tierpsychologie*, *55*(2), 139–160.

Pilley, J. W., & Hinzmann, H. (2013). *Chaser: Unlocking the genius of the dog who knows a thousand words*. Houghton Mifflin Harcourt.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 275.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of memory and language*, *57*(1), 1–23.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, *1*(3), 181–210.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic english grammatical structures: A corpus analysis. *Journal of memory and language*, *57*(3), 348–379.

Sag, I. A. (2010). Feature geometry and predictions of locality. *Features: Perspectives on a key notion in linguistics*, 236–271.

Sag, I. A., Wasow, T., Bender, E. M., & Sag, I. A. (1999). *Syntactic theory: A formal introduction* (Vol. 92). Center for the Study of Language and Information Stanford, CA.

Saldana, C., Oseki, Y., & Culbertson, J. (2019). Do cross-linguistic patterns of morpheme order reflect a cognitive bias? In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting fo the cognitive science society* (pp. 994–1000). Cognitive Science Society.

Saldana, C., Smith, K., Kirby, S., & Culbertson, J. (2018). Is regularisation uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation in morphology and word order.

Savage-Rumbaugh, E. S., Murphy, J., Sevcik, R. A., Brakke, K. E., Williams, S. L., Rumbaugh, D. M., & Bates, E. (1993). Language comprehension in ape and child. *Monographs of the society for research in child development*, i–252.

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in nicaragua. *Science*, *305*(5691), 1779–1782.

Steedman, M. (1993). Categorial grammar. *Lingua*, *90*(3), 221–258.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language learning and development*, *3*(1), 1–42.

Tily, H., Frank, M., & Jaeger, F. (2011). The learnability of constructed languages reflects typological patterns. In *Proceedings of the 33rd annual meeting of the cognitive science society.* Cognitive Science Society.

Tomasello, M. (2000a). First steps toward a usage-based theory of language acquisition. *Cognitive linguistics*, *11*(1/2), 61–82.

Tomasello, M. (2000b). The item-based nature of children's early syntactic development. *Trends in cognitive sciences*, *4*(4), 156–163.

Tomasello, M. (2009). The usage-based theory of language acquisition. In *The cambridge handbook of child language* (pp. 69–87). Cambridge Univ. Press.

Truswell, R. (2017). Dendrophobia in bonobo comprehension of spoken english. *Mind & Language*, *32*(4), 395–415.

Wagers, M. W., Borja, M. F., & Chung, S. (2018). Grammatical licensing and relative clause parsing in a flexible word-order language. *Cognition*, *178*, 207–221.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, *58*(2), 250–271.

Wu, T. (2011). The syntax of prenominal relative clauses: A typological study. *Linguistic Typology*, *15*(3), 569–623.

Yip, V., & Matthews, S. (2007). Relative clauses in cantonese-english bilingual children: Typological challenges and processing motivations. *Studies in Second Language Acquisition*, *29*(2), 277–300.

Appendix A

Supplementary Analysis 1

Table A1 reports the results of analyses in which correct productions on untrained trials were modeled as a function of GROUP and TRAINED TRIAL ACCURACY, or individuals' mean performance accuracy on trained trials. These analyses are similar to the "subset" analyses reported throughout the paper, but rather than a categorical independent variable for performance on TRAINED trials ($< 50\%$ vs. $> 50\%$), here performance on TRAINED trials was treated as a continuous independent variable. We excluded data from participants with 0% accuracy on trained trials here, so that participants who simply were not trying (or were using a strategy like "just speak English") could not drive an effect in this analysis.

The subset models tested the hypothesis that the $> 50\%$ group's performance on untrained trials was better than a liberal estimate of the educated guessing rate (i.e., the $< 50\%$ group's performance on untrained trials). However, as an anonymous reviewer pointed out, 50% is an arbitrary threshold that might have had unintended consequences for the results and/or their interpretation. Furthermore, while the idea that acquisition of the syntactic representations is categorical is a helpful caricature, the reality is likely to be much more complex. We can better capture this complexity with a continuous variable.

The present analyses test the same idea, but in a way that is sensitive to the possibility of graded differences in acquisition. Here, a significant main effect of TRAINED TRIAL ACCURACY indicates that as participants' mean accuracy on TRAINED trials increased, so did their performance on UNTRAINED trials. Consistent with what is reported for the "subset" analyses, this effect was significant for Experiments 1a, 2, and 3, and marginally significant for Experiment 1b. The significant interaction in Experiment 1b reflects the fact that the DORC-ONLY group showed a stronger positive relationship between performance on TRAINED and UNTRAINED trials than the SRC-ONLY group (for whom the main effect was only marginally significant).

These results are similar to those reported in the main body of the paper (Tables

4, 5, 10, and 15), and thus are consistent with our interpretation of the data.

Table A1

Results of predicting performance on untrained trials with a continuous variable reflecting participants' mean accuracy on trained trials.

	Model results			
	β	z	p	
Experiment 1a				
Intercept	−9.957	−3.628	< .001	***
GROUP: DORC-ONLY	2.842	0.970	.332	
TRAINED TRIAL ACCURACY	7.191	2.115	.034	*
Interaction	3.424	0.814	.416	
Experiment 1b				
Intercept	−3.950	−3.811	< .001	***
GROUP: DORC-ONLY	−2.378	−1.696	.090	.
TRAINED TRIAL ACCURACY	2.804	1.825	.068	.
Interaction	5.211	2.272	.023	*
Experiment 2				
Intercept	−3.726	−4.696	< .001	***
GROUP: SRC+IORC	−1.817	−1.359	.174	
GROUP: DORC+IORC	−4.585	−1.527	.127	
TRAINED TRIAL ACCURACY	5.821	4.878	< .001	***
Interaction: SRC+IORC, UNTRAINED	0.235	0.122	.903	
Interaction: DORC+IORC, UNTRAINED	−0.397	−0.103	.918	
Experiment 3				
Intercept	−5.625	−3.268	.005	**
GROUP: DORC+IORC	−7.241	−0.598	.099	.
TRAINED TRIAL ACCURACY	9.613	2.963	.001	**
Interaction	3.439	−1.309	.554	

Appendix B

Supplementary Analysis 2

As one reviewer points out, languages with nominal case marking tend to also allow scrambling, or multiple constituent orders. If scrambling is a consequent of case marking, the reviewer cleverly suggests, then in Experiment 3 it might be more accurate to code as correct any constituent order inside the relative clause. For instance, for a subject relative clause, we might code as correct not only *indirect object, direct object, verb, head noun* word orders, but also *direct object, indirect object, verb, head noun*; *indirect object, verb, direct object, head noun*; and so on.

We therefore recoded the data from Experiment 3 so as to count as correct any response that was counted as incorrect in our previous analysis for reasons of relative clause-internal word order. Responses with other errors were still coded as incorrect (e.g., if the head noun appeared inside of the relative clause or if the case marking was incorrect).

Table B1 gives the proportion of correct response by group and trial type; Table B2 gives the results of the same model reported in Table 15 (Experiment 3 results) but using this new dependent variable. Only two cells changed from what was reported above: the SRC+DORC group produced 19 more correct DORCs under this coding scheme (increasing from 45.6% accuracy to 53.2%), and the DORC+IORC group produced 7 more SRCs (increasing from 17.9% to 20.6%). The other cells remained the same, largely due to the fact that deviations from the expected word order were most often accompanied by errors like mistakes in case marking.

Table B1

Experiment 3: proportion correct responses by group and structure type, coding scrambled responses as correct.

	TRIAL TYPE		
GROUP	SRC	DORC	IORC
SRC+DORC	.468	.532*	.401
DORC+IORC	.206*	.421	.583

* These cells changed with respect to the numbers reported in Table 14.

Table B2

Experiment 3 results, allowing for scrambling.

	Model results			
	β	z	p	
Model 1: GROUP × TRIAL TYPE				
Intercept	0.082	0.122	.903	
GROUP: DORC+IORC	−0.167	−0.182	.856	
TRIAL TYPE: UNTRAINED	−1.943	−1.639	.101	
Interaction	−3.234	−1.929	.054	.

These proportions in Table B1 are numerically very similar to those reported in the original analysis (Table 14). Statistically, the results are similar, but what was a significant interaction in the main analysis (Table 15; p=.009) is now only marginal (p=.054). Given the overall similarity to our previous results, there appears to be no need to question the interpretation of this experiment.

Appendix C

Supplementary Analysis 3

Another reviewer points out that, by coding responses with, for instance, case-marking errors, as incorrect, we may have unintentionally introduced bias into our data. That is, if a participant did not learn the case marking rules, they may still have learned the target relative clause structure. If some training conditions made learning case marking harder than other training conditions, then these groups may have been unfairly penalized in our coding of the test data.

To determine whether anything hinged on this particular aspect of the coding scheme, we recoded the data from Experiments 2 and 3. We excluded any trials where the response was incorrect for reasons that did not bear on whether the relative clause was well-formed. This included trials with agreement errors, errors in the nominal domain (e.g., missing determiners (Experiment 2) or case marking errors (Experiment 3)), and errors in the matrix clause. The most common remaining errors were word order errors and, in Experiment 2, cases where the relative clause was not positioned correctly relative to the head noun and the head noun's determiner (e.g., "the [__ to the girl flowers gave] the boy" instead of "the [__ to the girl flowers gave] boy").

However, this coding scheme resulted in the exclusion of the majority of trials from the untrained conditions, but far fewer of those from the trained conditions. For our original analysis of Experiment 2, we analyzed data from 1280 untrained trials. With this new coding scheme, only 566 untrained trials remained. For Experiment 3, we originally analyzed 504 untrained trials, but here only 336 untrained trials remained. (The numbers of trained trials decreased as well, though not as dramatically.)

The consequence of this was that the proportion of correct responses in untrained conditions which remained became very high – in some cases higher than the proportion of correct responses on trained trials; see Tables C1 and C3. The results of the main analyses applied to these new dependent variables are reported in Tables C2 and C4.

The results of these analyses differ from those in reported above in Tables 10 and 15, but not in a way that changes our interpretation of the data. (Indeed, the higher

Table C1

Experiment 2: proportion correct responses by group and trial type, excluding trials with errors that do not bear on the main hypothesis.

	TRIAL TYPE	
GROUP	Trained	Untrained
SRC+DORC	.843	.921
SRC+IORC	.865	.444
DORC+IORC	.955	.153

Table C2

Experiment 2 results, excluding trials with errors that do not bear on the main hypothesis.

	Model results			
	β	z	p	
Model 1: GROUP × TRIAL TYPE				
Intercept	2.702	4.378	< .001	***
GROUP: SRC+IORC	0.072	0.084	.933	
GROUP: DORC+IORC	2.493	2.771	.006	**
TRIAL TYPE: UNTRAINED	0.621	1.687	.092	.
Interaction: SRC+IORC, UNTRAINED	−4.073	−8.354	< .001	***
Interaction: DORC+IORC, UNTRAINED	−9.720	−9.488	< .001	***

Table C3

Experiment 3: proportion correct responses by group and trial type, excluding trials with errors that do not bear on the main hypothesis.

	TRIAL TYPE	
GROUP	Trained	Untrained
SRC+DORC	.630	.656
DORC+IORC	.608	.247

Table C4

Experiment 3 results.

	Model results			
	β	z	p	
Model 1: GROUP × TRIAL TYPE				
Intercept	1.865	1.433	.152	
GROUP: DORC+IORC	−0.436	−0.259	.796	
TRIAL TYPE: UNTRAINED	1.761	0.843	.399	
Interaction	−7.628	−1.880	.060	.

performance on untrained structures here is perhaps stronger support for the idea that participants acquired a single, general representation of relative clauses.) Similarly, the pattern of learning across training groups remained the same: in both expeirments, the SRC+DORC groups generalized to untrained structures more than the DORC+IORC groups, and in Expeirment 2, the SRC+IORC group generalized less than the SRC+DORC group but more than the DORC+IORC group.

The fact that this coding scheme resulted in cases where untrained structures were produced correctly more often than trained structures suggests that the coding scheme reported in the main body of the manuscript probably more closely reflects what we aim to measure: whether participants used a syntactic representation of the untrained structure.