Language (2026), 1–42
doi:10.1017/S0097850726000032

Linguistic Society of America

GENERAL RESEARCH ARTICLE

Mind the gap: Learning the surface forms of movement dependencies

Laurel Perkins[1], Naomi H. Feldman[2,3], and Jeffrey Lidz[2]

[1]Department of Linguistics, University of California Los Angeles, United States
[2]Department of Linguistics, University of Maryland College Park, United States
[3]Institute for Advanced Computer Studies, University of Maryland College Park, United States

Corresponding author: Laurel Perkins; Email: perkinsl@ucla.edu

Received: 01 June 2024; **Revised:** 09 January 2026; **Accepted:** 13 January 2026

Keywords: language acquisition; computational modeling; statistical learning; nonadjacent dependencies; movement

Abstract

In acquiring a syntax, children must detect evidence for abstract structural dependencies that can be realized in variable ways in the surface forms of sentences. In *What did David fix?*, learners must identify a nonlocal relation between a fronted object of the verb (*what*) and the phonologically null 'gap' in canonical direct object position after the verb, where it is thematically interpreted. How do learners identify a nonadjacent dependency between an expression and something that has no overt phonological form? We propose that identifying abstract syntactic dependencies requires statistical inference over both overt linguistic material and unsatisfied grammatical expectations: noticing when a predicted argument for a verb is unexpectedly missing may serve as evidence for the gap of an argument movement dependency. We provide computational support for this hypothesis. We develop a learner that uses predicted but unexpectedly missing objects of verbs to identify possible gaps of object movement, and identifies which surface morphosyntactic properties of sentences are correlated with these possible movement gaps. We find that it is in principle possible for a learner using this mechanism to identify the majority of sentences with object movement in child-directed English, and that prior knowledge of which verbs require objects provides an important guide for identifying which surface distributions characterize object movement. This provides a computational account for why verb argument-structure knowledge developmentally precedes the acquisition of movement in a language like English. More broadly, these findings illustrate how statistical learning and learning from violated expectations can be combined to novel effect in the domain of language acquisition.

1. Introduction

In acquiring a syntax for their native language, children infer a system that specifies ways of combining expressions in hierarchical structures and defines dependencies over those structures. These dependencies encode abstract grammatical relations, determined not by the specific form of any particular expression, but rather by the syntactic properties of expressions and their structural positions relative to each other.

For instance, the predicate-argument dependency between a verb and its direct object is established through a particular structural configuration in 1a, and it is the same regardless of the particular verb or the particular object noun phrase (in bold). And whereas in English this dependency is often established locally, between two adjacent expressions, the same abstract dependency can also be established nonlocally, across potentially large amounts of linguistic material. In each of the sentences in 1b–d, a

fronted phrase bears the same object relation to the verb *fix* as does the corresponding phrase (*a toy*) in 1a, despite appearing in a nonadjacent position.

(1) a. David is fixing **a toy**. Amy is buying **a plane ticket**.
b. **What** did David fix?
c. **What** did the girl who we saw at the park say that David fixed?
d. I found **the toy** that David fixed.

These examples show us that syntactic dependencies are highly abstract in relation to the specific forms that express them. The same verb-object dependency can be satisfied by phrases with very different surface forms, appearing in very different positions in a sentence. And these dependencies take still different forms in other languages. This tension between the abstract nature of syntactic dependencies and the variability of surface forms that realize them presents a challenge for theories of how this central domain of syntax is acquired (Chomsky 1965, 1980, Fodor 1998, Lidz & Gagliardi 2015, Pinker 1984, Valian 1990). How do language learners come to identify abstract structural relations in the face of such great variety in surface expression?

Prior accounts of dependency acquisition have largely focused on dependencies that are morphologically marked, such as the relation between the auxiliary verb *is* and the *-ing* form of the verb in 1a. Young children show awareness of the cooccurrence patterns of nonadjacent sounds and morphemes in their input, statistical sensitivities that may allow them to discover morphosyntactic dependencies at early ages (Gómez 2002, Gómez & Maye 2005, Höhle et al. 2006, Nazzi et al. 2011, Santelmann & Jusczyk 1998, Tincoff et al. 2000, van Heugten & Shi 2010). But this represents only a narrow corner of the dependencies that learners must acquire. Here, we turn our attention to the sorts of dependencies illustrated in 1b–d, in which an object is moved from its canonical position after the verb.[1] The abstract nature of movement dependencies poses a challenging learning problem. Identifying that the same verb-object dependency is present in 1a and 1b–d requires tracking the cooccurrences not only of specific surface forms, but also of abstract syntactic categories and positions. Learners must become aware that a fronted noun phrase is standing in a nonlocal relation to something that has no overt phonological form: the 'gap' associated with the verb, in canonical direct object position, where it is thematically interpreted.

In this article, we argue that identifying abstract syntactic dependencies requires statistical inference over both overt and hidden grammatical structure. We pursue the hypothesis, consistent with a broader literature on the role of expectation violation in development (Denison & Xu 2012, Kouider et al. 2015, Stahl & Feigenson 2015, 2017, Téglás et al. 2011), that children learn from unsatisfied grammatical predictions. Our case study is the role of verb argument-structure knowledge in the acquisition of argument movement. In their second year of life, children begin to identify subjects and objects in their canonical positions, and to learn which verbs require objects (Fisher et al. 2019, Jin & Fisher 2014, Lidz et al. 2017, White & Lidz 2022, Yuan et al. 2012). Movement dependencies are acquired only after local argument-structure knowledge has emerged (Gagliardi et al. 2016, Perkins & Lidz 2020, 2021). This developmental trajectory points toward a particular learning mechanism: knowledge of local argument dependencies may help learners identify when arguments have been moved. If children notice when a predicted argument for a verb is missing in its expected position, this may compel them to search for that argument nonlocally, and thereby learn the morphosyntactic footprints of particular movement dependencies in their language (Gagliardi et al. 2016, Perkins 2019, Perkins & Lidz 2020, Stromswold 1995).

We provide computational support for this proposal. We develop a learner that identifies which surface morphosyntactic properties of sentences are correlated with expected but missing direct objects of verbs. In simulations on child-directed English, our model successfully identifies the majority of sentences with object movement in its input. Moreover, we show that prior argument-structure knowledge plays a

[1] Here, 'move' simply means that the relation between the object and the verb is established nonlocally. Any syntactic theory needs to account for the fact that the same dependency can be satisfied both locally and nonlocally. We use 'move' as a theory-neutral term for this phenomenon.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

substantial role in the success of this distributional learning mechanism: knowledge of which verbs require objects provides an important guide for identifying which surface distributions characterize object movement. These findings provide insight into how learning from expected grammatical structure can work in concert with statistical learning to enable syntactic dependency acquisition in early development.

2. Acquiring nonlocal syntactic dependencies

A large body of literature finds that sensitivity to dependencies between nonadjacent sounds and morphemes develops in an infant's second year of life (Gómez 2002, Gómez & Maye 2005, Höhle et al. 2006, Nazzi et al. 2011, Santelmann & Jusczyk 1998, Tincoff et al. 2000, van Heugten & Shi 2010). For instance, Santelmann & Jusczyk 1998 showed that eighteen-month-old English learners are aware of the dependency between *is* and *-ing* in sentences like *Everybody is baking bread.* Because these types of nonadjacent dependencies are morphologically marked, they leave detectable evidence on the surface forms of sentences that learners hear. That is, to identify that there is a dependency between *is* and *-ing*, learners need only notice that these sounds cooccur in their input with unusual regularity—although this still leaves open the question of how learners identify that this surface-level cooccurrence is marking a particular grammatical dependency, namely the relation between the auxiliary *be* and a verb in the progressive aspect (Höhle et al. 2006, Nazzi et al. 2011, Santelmann & Jusczyk 1998, Tincoff et al. 2000).

Other types of nonlocal syntactic dependencies, such as the argument movement dependencies in WH-questions, have received much less attention in prior work. These also pose a more substantial learning challenge. English WH-phrases have different surface forms from clause arguments in their canonical positions and have different distributions: they overwhelmingly occur clause-initially. Therefore, recognizing that the same verb-object dependency is present in the WH-question in 1b and in the basic transitive clause in 1a requires abstracting away from these surface properties. Infants cannot merely track the cooccurrences of specific sounds or lexical items; they must represent the dependency abstractly, as an instance of the same dependency that is typically established locally between a verb and its direct object.

Prior experimental work has found that infants as young as fifteen months sometimes respond appropriately to WH-questions (Gagliardi et al. 2016, Perkins & Lidz 2020, Seidl et al. 2003). But Gagliardi et al. (2016) and Perkins and Lidz (2020) argue that infants' success on these tasks may reflect an interpretive heuristic based on knowledge of local argument dependencies in combination with pragmatic reasoning, rather than syntactic representations of the nonlocal dependencies in these questions. This argument is motivated by earlier findings that children at fifteen to sixteen months show sensitivity to lexical and clause transitivity (Jin & Fisher 2014, Lidz et al. 2017). Learners at this age are beginning to identify which verbs require direct objects (Lidz et al. 2017), and in the following months they gain facility in using this knowledge to predict upcoming direct objects during on-line sentence processing (Hirzel et al. 2020, Lidz et al. 2017, White & Lidz 2022). Infants in this age range also use subjects and objects to draw inferences about verb meaning, interpreting verbs with both subjects and objects as labels for causal events (Jin & Fisher 2014). This early knowledge of local subject and object dependencies may lead to the appearance of WH-question comprehension in prior preferential looking tasks, even without representing WH-dependencies syntactically. Such tasks typically presented infants with WH-questions with transitive verbs, such as *Which dog did the cat bump?*, in the context of events in which, for example, a dog bumps a cat, and the cat bumps a different dog. A fifteen-month-old who can identify that *the cat* is the subject in this question, and who knows that *bump* typically requires a direct object, may be inclined on the basis of that knowledge to look at an individual who got bumped by a cat—appearing to understand the question without necessarily representing *which dog* as a nonlocal object of the verb. In support of this account, Perkins & Lidz 2020 found that fifteen-month-olds' performance on this task depended on their vocabulary, a likely index of their verb knowledge.

Perkins & Lidz 2021 provided a more rigorous test of WH-dependency representations by asking when infants register the complementarity between a local direct object and an object WH-phrase. If infants represent the WH-phrase in a sentence like 1b as expressing the same grammatical relation as the local

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

direct object in 1a, then they should be aware that the WH-phrase cannot cooccur with a local object: **What did David fix a toy?* is ungrammatical. In a listening preference task, infants were presented with both WH-questions and basic declarative clauses with transitive verbs, with and without local direct objects. Eighteen-month-olds listened longer to basic declarative sentences with local objects versus without (e.g. *A dog! The cat should bump him!* > **A dog! The cat should bump!*), but displayed the opposite pattern of preference for WH-questions (e.g. *Which dog should the cat bump?* > **Which dog should the cat bump him?*). That is, eighteen-month-olds showed a consistent preference for grammatical sentences of each type. However, fourteen- and fifteen-month-olds did not differentiate between these sentence types. These results suggest that infants represent the WH-phrase as a nonlocal object of the verb at eighteen months, but not before.

2.1. Learning mechanisms

The experimental results surveyed above point toward the following developmental trajectory. Basic verb argument-structure knowledge appears to develop early, at fifteen to sixteen months for English learners, and emerges before infants identify moved arguments, such as those in WH-questions. What learning mechanisms might allow learners to identify these nonlocal argument dependencies in their input? This is not a trivial task. Movement dependencies are not always marked with consistent morphology: for instance, English WH-phrases take a variety of different forms. The class of WH-elements in any language will distribute in specific ways in the surface forms of sentences: for instance, English WH-words are clause-initial and frequently occur in questions. However, even if a learner can identify a word class with these particular surface distributional properties, it does not necessarily follow that these are WH-elements. Many languages have question particles that can appear at sentence boundaries in both WH- and polar questions. An example is the particle *la* in Tz'utujil Mayan, as in 2. A Tz'utujil learner needs a way to tell that *la* is a question particle and not a WH-word, and conversely an English learner needs a way to tell that *what* is a WH-word and not a question particle.

(2) Tz'utujil Mayan (Dayley 1981)
La xwari ja ch'uuch'?
Q slept the baby
'Did the baby sleep?'

Moreover, in many languages, WH-phrases do not appear clause-initially. In WH-in-situ languages like Chinese, Japanese, and Korean, WH-phrases are pronounced in their thematic position local to the verb, but still take interrogative scope in a higher clausal position, as in 3. Learners of these languages must identify when an expression is in a nonlocal WH-dependency with a higher clausal node, even when it has not overtly moved to this position.[2]

(3) Mandarin Chinese (Cheng 2003)
Hufei mai-le shenme?
Hufei buy-PRF what
'What did Hufei buy?'

Thus, in order to identify WH-dependencies in their language, children must solve multiple problems. They need to learn whether their language fronts WH-phrases, and if so, which surface forms signal that this movement has occurred. In a language with WH-fronting, they also must identify the thematic position where the WH-phrase should be interpreted in relation to the verb. As noted above, WH-in-situ poses a

[2] On many accounts, this scope relation is established through covert movement (e.g. Aoun et al. 1981, Huang 1982). Other nonmovement accounts of WH-in-situ include binding by a covert operator (Reinhart 1998), with some proposing different WH-in-situ representations across different languages (Cole & Hermon 1994). See Cheng 2003 for an overview.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

different learning problem from WH-fronting: in WH-fronting, a WH-phrase is pronounced in the position where it takes interrogative scope, and learners must identify a nonlocal dependency with its thematic position, whereas in WH-in-situ, the WH-phrase is pronounced in its thematic position, and learners must identify a nonlocal dependency with its scope position. We focus here on the problem posed by WH-fronting, but return to consider WH-in-situ in the general discussion in Section 5.

In English, surface signals for WH-movement include not only WH-words, but also a variety of other reflexes of movement, such as prosodic marking and, in questions where the moved constituent is not a subject, subject-auxiliary inversion and *do*-support. Mature speakers of a language make efficient use of these signals in sentence processing to identify moved arguments and predict upcoming 'gaps' where they should be interpreted (Aoshima et al. 2004, Crain & Fodor 1985, Frazier & Clifton 1989, Frazier & Flores d'Arcais 1989, Sussman & Sedivy 2003, Traxler & Pickering 1996). But children must first learn these signals in order to use them in parsing WH-dependencies. In languages like English, identifying the tails of these dependencies is particularly challenging, because the thematic positions of moved elements are phonologically null. How do learners identify a nonadjacent dependency where only one element appears overtly?

One possible piece of the puzzle comes from the literature on 'expectation violation' or 'error-driven learning' in other areas of cognitive development. A large body of work finds that infants use knowledge about the physical and social properties of objects and agents, alone or in combination with learned statistical contingencies, to make predictions about upcoming events (Denison & Xu 2012, Kouider et al. 2015, Stahl & Feigenson 2015, 2017, Téglás et al. 2011). Violations of these predictions may provide valuable opportunities for learning (Stahl & Feigenson 2015, 2017). For instance, an experiment in Stahl & Feigenson 2015 presented eleven-month-olds with events that either conformed with or violated object solidity. In one such event, a ball rolled down a ramp toward a solid wall, stopping behind an occluder. When the occluder was lifted, one group of infants saw that the ball had been stopped by the wall, while a second group of infants saw that the ball had apparently passed through the wall, violating their predictions about object solidity. After this event, both groups of infants were tested on their ability to map a novel property (e.g. squeaking) to the previously observed toy. Infants who had observed the prediction-violating event showed significantly greater learning than infants who had not. In a further experiment, infants who viewed these events were then given a choice to explore the ball or a novel object. Infants who had viewed the prediction-violating event chose to explore the ball more than infants who had not. Moreover, their exploration was consistent with testing the object's solidity properties: they banged the ball against the table to a greater extent than infants who had seen a different event type. These results suggest that even very young learners are sensitive to inconsistency between their own predictions and observed events, and when they observe a situation where their predictions are violated, they exploit this opportunity to learn, explore, and test hypotheses about the potential cause of that violation.

We pursue the hypothesis that a similar form of expectation violation may underlie infants' discovery of argument movement dependencies in languages like English. Here, it is not predictions about physical events that drive learning, but rather predictions about grammatical structure. On this hypothesis, verb argument-structure knowledge developmentally precedes argument movement acquisition because the former provides the basis for generating structural predictions—specifically, predictions about upcoming arguments of verbs. When infants encounter a case where an expected argument does not appear in its local position, they exploit this expectation violation to learn about the cause of the locally missing argument, scaffolding their identification of movement dependencies (Gagliardi et al. 2016, Perkins 2019, Perkins & Lidz 2020, Stromswold 1995). For example, learners who know that a verb like *fix* requires a direct object might register that it is unexpectedly missing after the verb in a question like *What did David fix?*. This unsatisfied structural prediction may provide the basis of inferring the tail of a nonlocal argument dependency—a 'gap' of argument movement—even though it is silent. And it may compel learners to search the rest of the sentence for the cause of the missing argument, eventually identifying that another expression in the sentence (*what*) is satisfying the verb's transitivity requirement nonlocally. This would allow them both to assign an appropriate parse to the sentence and to begin to learn how various

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

types of nonlocal dependencies are realized: that is, that this question contains a WH-dependency, which is marked in English by various surface signals, such as *what*, *do*-support, and subject-auxiliary inversion.

In sum, we propose that the process of acquiring nonlocal dependencies follows three logically independent steps, which we together call **Gap-Driven Learning** (Perkins 2019, Perkins & Lidz 2020):

(i) using knowledge of verb argument structure to detect argument gaps: predicted arguments that are unexpectedly missing in their local positions;
(ii) identifying what surface forms are correlated with these argument gaps; and
(iii) inferring what types of syntactic dependencies are responsible for those correlations.

Here, we investigate the gap-driven learning hypothesis specifically in the domain of direct object gaps. This decision is motivated by empirical evidence for early knowledge of verb transitivity (Jin & Fisher 2014, Lidz et al. 2017), making it plausible that direct object gaps are the type of argument gap that learners may be able to detect readily at the relevant stage of development. But how this knowledge is in place by this age raises its own learning problem, which must be addressed in order for gap-driven learning to be possible. Before children can identify when arguments have been moved, they cannot identify all instances of direct objects in sentences containing transitive verbs. How, then, do they arrive at the appropriate expectations that some verbs obligatorily require objects, such that they will be surprised when those objects are missing? Perkins et al. (2022) investigate this question computationally and show that it is feasible for children to find their way around this learning problem. The learner in Perkins et al. 2022 assumes that it occasionally represents sentences erroneously and learns what portion of its input representations to treat as signal versus noise for the purpose of learning verb transitivity. When tested on the distributions of direct objects that a child at this age could identify in child-directed English, the model learned how to filter its data to correctly assign transitivity properties to the majority of the most frequent verbs in its input. This tells us that it is in principle possible for children to identify verb transitivity without accurately parsing argument movement, thereby providing a way for gap-driven learning to get started.

In this article, we present a computational model that instantiates the first two steps of learning under the gap-driven learning hypothesis. The learner builds off of the model in Perkins et al. 2022, using the approximate verb transitivity knowledge identified by that learner. Our model tracks statistical regularities in the surface morphosyntactic features of sentences in order to identify clusters of sentences that share distributional properties. At the same time, it tracks when its expectations of upcoming direct objects are violated, in order to infer which clusters of properties are correlated with potential direct object gaps. When tested on child-directed speech, we find that the model identifies the large majority of sentences with object movement. Furthermore, we show that prior knowledge of verb transitivity, even if rough and approximate, is important for this distributional learning process to be successful. The learner performs better if it uses transitivity knowledge to infer likely object gaps, rather than clustering sentences on the basis of their overt surface features alone. These findings demonstrate that a learner could in principle identify object movement dependencies in English by using unsatisfied structural predictions to guide distributional learning. As verb transitivity knowledge forms the basis for generating these structural predictions, this provides an account for the empirically attested order of argument structure and argument movement acquisition in early development.

3. Model

We present a Bayesian model that simultaneously tracks the statistical distributions of surface morphosyntactic features in sentences and applies its knowledge of verb transitivity in order to infer which distributional properties are correlated with locally missing direct objects. This distributional learning takes the form of categorization: the learner infers 'categories' of sentences according to their feature distributions, and infers which sentence categories likely contain direct object gaps. When the learner sees a sentence that violates its expectations about verb transitivity, the learner infers that this sentence contains a direct object gap and that all other sentences in the distributionally defined category do so as well. This

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

allows the learner to generalize across sentences that share similar surface features, and to infer which of those shared features signal object movement dependencies.

This distributional learning mechanism follows prior computational work that has proposed similar mechanisms for the acquisition of phonetic categories in infancy and for category learning domain-generally (Anderson & Matessa 1990, Feldman et al. 2013, Maye et al. 2002, McMurray et al. 2009, Sanborn et al. 2010). Similar to these previous models, the current account envisions the learning task as requiring two simultaneous inferences: discovering the underlying system of categories that give rise to distributions of surface features that a learner observes, and identifying which observations belong to which category. However, it departs from previous literature by envisioning this categorization process as merely a means to an end. Whereas the phonetic learning literature has traditionally assumed that there is a set of phonetic categories to be acquired (but see Feldman et al. 2021), here we do not assume that adult grammars necessarily represent 'categories' of sentences in any meaningful way. Instead, the categories inferred by this learner are an intermediate step of learning: they enable further inference about the underlying properties of sentences that are formally similar. When the learner infers that one sentence in a category likely contains an object gap, it then infers that this property holds of other sentences in the category as well. In doing so, it identifies which surface features are correlated with object gaps and therefore may be the footprints of movement.

Our computational approach falls under the paradigm of Bayesian cognitive modeling. A cognitive model formalizes a hypothesis about the knowledge that a learner brings to a particular learning task (the learner's hypothesis space, containing assumptions about how its data are generated), along with the mechanisms that a learner uses to update that knowledge on the basis of new data. Bayesian approaches characterize learners' beliefs as probability distributions over hypotheses, which are updated using rational probabilistic inference: the **posterior** probability of a hypothesis given observed data is calculated by combining the learner's **prior** beliefs with the **likelihood** of the data under each hypothesis. The learner that we present in the current work is a Bayesian model that is 'nonparametric' in the sense that the size of its parameters (the number of latent sentence categories to be acquired) is unknown in advance. The approach taken in Bayesian cognitive modeling differs from the statistical approach of hypothesis testing through Bayesian regression: in the cognitive modeling paradigm, the model itself is the hypothesis being tested, rather than a tool for assessing which of several hypotheses provides the best fit for data. Such a model can take many different forms, depending on the theoretical assumptions of the modeler, and typically assumes a complex, nonlinear relationship between variables and data. The current approach uses some of the same techniques from the machine learning literature, but differs from supervised machine learning in that the model is not fit on the learning objective that it is tested on, so its data need not be split between training and test sets to avoid overfitting. For a detailed tutorial introduction on Bayesian cognitive methods and further examples of how this paradigm has been applied, see Griffiths et al. 2024 and the citations therein.

Following a rich tradition in the language acquisition literature (e.g. Abend et al. 2017, Alishahi & Stevenson 2008, Berwick 1985, Dillon et al. 2013, Elman 1990, Frank et al. 2009, Goldwater et al. 2009, Pearl & Sprouse 2019, Perfors et al. 2010, Perfors et al. 2011, Perkins et al. 2022, Sakas & Fodor 2001, 2012, Vallabha et al. 2007, Wexler & Culicover 1980, Yang 2002), our model is framed at Marr's (1982) computational level. We aim to characterize a particular type of mental computation that could give rise to successful learning given the information available in children's data and a set of hypotheses about their knowledge at the relevant developmental stage. This model therefore represents an idealization of learners' actual inference processes, but an idealization that is nonetheless grounded in empirical data about their grammatical knowledge and representational abilities in development, described in more detail below. It also provides a measure of how much information is available in the child's representation of the input (at a particular stage of development) to support the hypothesized inferences. The results of our simulations open the door for further algorithmic questions concerning learners' abilities to access and use the information available in their environment, and whether their learning processes resemble this idealized mechanism.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

In this section, we (i) specify the generative model, encoding the learner's assumptions about how its observations of sentence features are generated, and (ii) specify how the learner jointly infers sentence categories and object gaps, given its data and its knowledge of verb transitivity. The following sections present simulations demonstrating that this joint inference allows the learner to successfully identify features that characterize object movement dependencies in English, when tested on child-directed speech.

3.1. Generative model

The data that our learner observes consists of the morphosyntactic features of sentences containing transitive, intransitive, or alternating verbs. The learner builds off of a first step of learning modeled in Perkins et al. 2022, which shows how some initial knowledge of verb transitivity properties might be acquired before a child can identify moved objects. That learner assumed that there are three transitivity categories to be identified—transitive verbs that require direct objects, intransitive verbs that disallow them, and alternating verbs that optionally allow them—and assigned verbs in its input to these three categories based on their distributions with direct objects in canonical postverbal positions, which English-learning infants can identify prior to eighteen months (Gagliardi et al. 2016, Hirsh-Pasek & Golinkoff 1996, Jin & Fisher 2014, Lidz et al. 2017, Perkins & Lidz 2020, Seidl et al. 2003). These initial transitivity assignments are imperfect, modeling the realistic assumption that a child's knowledge of verb transitivity is likely to be approximate at this stage of development.

The current learner now assumes that there are two reasons why it might observe canonical direct objects or no direct objects after the verbs in the sentences that it observes. On the one hand, the transitivity of that verb determines whether it should always, never, or sometimes occur with a direct object. On the other hand, there may be a separate grammatical process, such as argument movement, that results in an apparent transitivity violation. The learner assumes that these transitivity violations are governed by latent 'categories' of sentences with shared grammatical properties. Each category has a particular parameter governing whether it produces object gaps: if it does, then observations of canonical direct objects in that category may no longer reflect the transitivity properties of these verbs, but may instead be due to other grammatical properties that produce 'nonbasic' word orders. These properties also give rise to the distributions of other morphosyntactic features of sentences in a particular category.

For instance, the learner might identify that a sentence like *What did David fix?* belongs to a category of other sentences that have object gaps and also tend to be questions with subject-auxiliary inversion, a form of *do*, and an unknown functional element sentence-initially (e.g. *what*). On the other hand, the learner might identify that a sentence like *Your toy got broken* belongs to another category of sentences that also have object gaps, but different morphosyntactic features: here, a form of *get* and the verbal suffix *-en.* The distributional features of the first sentence category are the footprints of object WH-questions in English; the features of the second category are the footprints of *get*-passives.

The learner does not know ahead of time how many sentence categories there will be or what the properties of those categories are. Using the distributions of direct objects and the other observed sentence features in its data, the learner infers what categories of sentences are present, what their distributional properties are, and which categories produce object gaps. This allows the learner to identify specific clusters of morphosyntactic features that are correlated with object gaps in different clause types, which may be candidates for entering into nonlocal movement dependencies.

More formally, we provide the graphical model for the learner in Figure 1. A graphical model provides a visual representation of the process by which the learner assumes its data are generated. Circular nodes represent random variables, and arrows represent conditioning relationships between variables. Shaded nodes represent variables with observed/known values; unshaded nodes represent variables whose values are unknown and must be inferred. Rectangular 'plates' indicate when a portion of the model is repeated over a particular range, denoted by the superscript in the right corner. See Griffiths et al. 2024 for more information.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

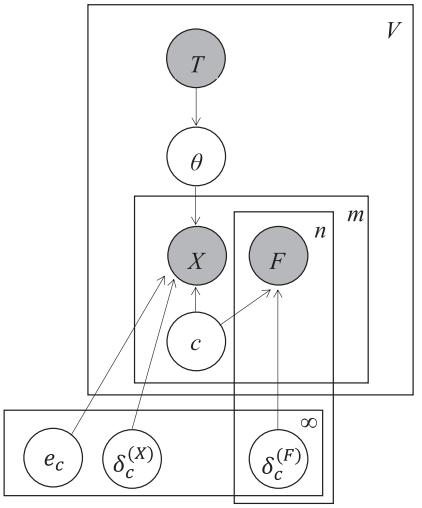

Figure 1. *Graphical model for Joint Inference Learner. Nodes correspond to random variables: the observed direct objects* X *and other features* F *in each sentence, the transitivity category* T *and rate of direct objects* θ *for each verb, the latent 'category'* c *of each sentence, the rate of direct objects* $\delta^{(X)}$ *and other sentence features* $\delta^{(F)}$ *produced by each category, and whether each category produces a transitivity violation* e. *Arrows denote conditioning relationships between variables.*

Observations of direct objects are formalized as the Bernoulli random variable X. This variable encodes direct object data for each of the m sentences containing each of the V verbs in the model's input, with a value of 1 if the sentence contains a direct object following the verb, and 0 if it does not. The model's observations of the other n relevant morphosyntactic features of the sentence are represented by the vector of Bernoulli random variables $\vec{F}$. Specific details of this feature set are discussed in the next section.

The direct object observations $X^{(v)}$ for a given verb v can be generated by two processes: the transitivity of verb v, represented by the variables T and θ in the upper half of the model, or the other grammatical properties of the category that the sentence belongs to, represented by the variables c, e, and $\delta^{(X)}$ in the lower half of the model. We describe each of these generative processes in turn.

In the upper part of the model, each observation $X^{(v)}$ of a direct object for a particular verb is conditioned on the parameter $\theta^{(v)}$, a continuous random variable that controls the probability that verb v will be used with a direct object. $\theta^{(v)}$ is conditioned on the variable $T^{(v)}$, a discrete random variable that can take on three values corresponding to transitive, intransitive, or alternating verbs. In order to model the hypothesis that learners are using prior knowledge of verb transitivity properties, we assume that the learner has approximate knowledge of these values of T for the set of verbs in the learner's data, as acquired by the model in Perkins et al. 2022. This means that the learner knows some of the values of θ as well. If verb v is fully transitive, then the learner assumes that $\theta^{(v)} = 1$: the verb should always occur with a direct object. If the verb is fully intransitive, then $\theta^{(v)} = 0$: the verb should never occur with a direct object. If the verb belongs to the alternating category of T, then $\theta^{(v)}$ takes an unknown value between 0 and 1 inclusive. The prior probability over θ in this case is a $Beta(\alpha,\beta)$ distribution, where the parameters α and β are counts of direct objects and no direct objects for verb v in sentence categories without transitivity violations, excluding the current category.

In the lower part of the model, each $X^{(v)}$ is conditioned on the discrete random variable c, defined for all positive integers, which represents the category that the sentence belongs to. These sentence categories also condition the other morphosyntactic features in the sentence, encoded in the vector $\vec{F}$. Each category c is assumed to reflect a particular set of underlying grammatical properties that give rise to the distributions of direct objects and other features of a sentence. The number and properties of these

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

categories are a priori unknown, and the learner infers the properties that will allow it to explain the distributions of features and direct objects that it observes. Returning to our earlier examples, the learner might infer a value of c that encodes English WH-object questions, giving high probability to sentence-initial function words (i.e. WH-words), subject-auxiliary inversion, forms of *do*, and direct object gaps. Another inferred value of c might encode English *get*-passives, giving high probability to direct object gaps, forms of *get*, and the *-en* verbal suffix. The prior probability over c is a Dirichlet process (Ferguson 1973), which gives a particular category prior probability proportional to the number of sentence observations already assigned to that category. This process also reserves a small nonzero probability for new categories, allowing the model to flexibly converge on the number of sentence categories that best explains the distributions in its data. By allowing the model to explore a potentially unbounded number of categories, this prior builds in the fewest possible assumptions about the number of categories required to explain the distributions in a given language or data set; however, this form of prior also biases the model to reuse categories whenever possible, and thus to keep the total number of categories small.[3] See Appendix A for details.

The random variables e, $\delta^{(X)}$, and $\delta^{(F)}$ represent the parameters of each of the sentence categories. The Bernoulli random variable e_c encodes whether a given category c produces transitivity violations. If $e_c = 0$, then the category does not produce transitivity violations, and all observations of a direct object in $X^{(v)}$ were generated by the transitivity properties of verb v. But if $e_c = 1$, then the category does produce transitivity violations, and the observations of direct objects $X^{(v)}$ were generated by a particular grammatical property of category c. The learner in Perkins et al. 2022 inferred that transitivity violations occurred approximately 19% of the time in sentences containing this same set of verbs in child-directed speech. In order to model the hypothesis that the current learner builds off of the knowledge gained in that previous stage of learning, our learner assumes that 19% is the prior probability that $e_c = 1$.[4]

The random variable $\delta_c^{(X)}$ represents the probability of observing a direct object in a category with transitivity violations—that is, whether the particular violation in that category produces object gaps, or whether it adds an apparent extra object that is not licensed by the verb. Intuitively, we can think of the probability that a sentence contains a direct object as depending on one of two biased coins. If $e_c = 0$ and the observation was generated by the verb's transitivity properties, then one biased coin is flipped and the sentence contains a direct object with probability $\theta^{(v)}$. But if $e_c = 1$ and the observation was generated by the grammatical properties of category c, then a different biased coin is flipped and the sentence contains a direct object with probability $\delta_c^{(X)}$. The parameter $\delta_c^{(X)}$ is assumed to have a uniform $Beta(1,1)$ prior distribution. This uniform prior means that it is equally likely a priori for a sentence category to create object gaps as it is to add extra objects. This form of prior builds in the fewest possible assumptions about the probability of observing a direct object versus an object gap within a sentence category. Analogous to $\delta_c^{(X)}$, the random variables in $\vec{\delta}_c^{(F)}$ represent the probabilities of observing the other morphosyntactic features in a given sentence category. Each $\delta_c^{(F)}$ is also assumed to have a uniform $Beta(1,1)$ prior distribution, meaning that all features are equally likely a priori to be present as they are to be absent; this likewise builds in the fewest possible assumptions about the distributions of features within sentence categories.

[3] Note that this is not intended as a claim about the number of sentence categories that a child's developing cognitive capacities can support. Following Anderson 1990, we frame the current model as a way to determine the solution that an ideal learner could arrive at from the input that a child is exposed to. Further work might extend this model to explore different hypotheses about limits on the number of categories that a child can feasibly entertain.

[4] The model in Perkins et al. 2022 differs from the current model in that it did not group sentences into categories. In the previous model, this parameter represented the probability of transitivity violations across sentences in the corpus. In the current model, this parameter represents the probability of transitivity violations across categories of sentences. These two parameters are not necessarily equivalent; they will be equivalent only if sentences are equally distributed among sentence categories. Although this assumption may not be borne out, it is adopted here as a simplifying assumption of the learner's prior, which can be overridden as the learner updates its hypotheses upon seeing data.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

3.2. Inference

The learner uses component-wise Gibbs sampling (Geman & Geman 1984) to jointly infer the category of each observed sentence (c) and whether each category contains transitivity violations (e). We first initialize values of c and e for each sentence. Then, for each sentence, we calculate a posterior probability distribution over new category assignments given the observed data in X and F, the known verb transitivity properties T, and the other sentence category assignments and properties. We resample new values of c for each sentence sequentially from this posterior probability distribution. Finally, we use the new category values to resample values of e for each category from its posterior probability distribution, given the other model parameters. This cycle is repeated over many iterations until the model converges to a stable distribution over c and e. Details of the initialization and sampling procedure are provided in Appendix A.

4. Simulations

We tested our learner on a data set of child-directed English. As described above, our model performs two steps of inference: it jointly categorizes sentences according to their surface feature distributions, and infers which sentence categories have direct object gaps. In order to evaluate its performance and assess the importance of each of these inference steps, we compared it to a baseline model that lacks one of these steps. The first baseline model uses verb transitivity knowledge to identify object gaps, but does not categorize sentences based on their feature distributions. The second baseline model categorizes sentences based on their feature distributions, but lacks verb transitivity knowledge and the ability to identify object gaps. We ask two primary questions: (i) how well can our learner identify instances of object movement in English, in comparison to these baselines? and (ii) how informative are the specific features of the model's categories for isolating movement dependencies from other grammatical processes?

4.1. Data

We prepared a data set from four parsed corpora in the CHILDES Treebank (Pearl & Sprouse 2013), which contains parse trees for child-directed English corpora on CHILDES (MacWhinney 2000). Details of these corpora are provided in Table 1. From these corpora, we selected sentences containing the verbs whose transitivity properties are known by our learner. Because a child's knowledge of verb transitivity is likely to be imperfect before eighteen months of age, we base our learner's knowledge on the transitivity classes inferred by the learner in Perkins et al. 2022, which provides a model of the previous stage of learning that our current model builds off of. We selected 18,503 sentences containing the verbs whose transitivity properties were inferred by the previous learner: these are the fifty most frequent transitive, intransitive, and alternating action verbs in these corpora. Because the previous learner assigned only 66% of these verbs to the correct transitivity category as specified in Perkins et al. 2022, this provides a noisy and imperfect source of knowledge

Corpus	# Children	Ages	# Words	# Utterances
Brown—Adam, Eve, & Sarah (Brown 1973)	3	1;6–5;1	391,848	87,473
Soderstrom (Soderstrom et al. 2008)	2	0;6–1;0	90,608	24,130
Suppes (Suppes 1974)	1	1;11–3;11	197,620	35,904
Valian (Valian 1991)	21	1;9–2;8	123,112	25,551

Table 1. *Corpora of child-directed speech.*

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

for the current learner.[5] Table 2 provides the frequencies of these verbs, along with the transitivity categories assumed by our model.

We conducted an automated search over the Treebank trees for overt direct objects following each verb, as well as the morphosyntactic features of each sentence that our model observes. We assume that the

Verb	Total	% Direct objects
Transitive		
feed	220	93%
fix	337	91%
pick	331	90%
bring	605	89%
drop	169	88%
throw	312	88%
hit	214	87%
lose	185	86%
close	166	85%
buy	358	84%
touch	183	84%
leave	356	83%
wash	195	83%
Alternating		
pull	331	81%
push	352	78%
open	342	77%
catch	185	76%
cut	263	75%
bite	191	73%
turn	485	72%
build	299	72%
knock	160	72%
hold	579	70%
read	509	69%
break	550	63%
drink	366	60%
wear	477	60%
eat	1,318	59%
sing	306	53%
blow	255	52%
draw	375	51%
move	238	47%
ride	281	41%
hang	151	35%
stick	192	29%
write	583	27%

Table 2. *Continued*

[5] The errors made by the learner in Perkins et al. 2022 were primarily in classifying verbs as deterministically transitive or intransitive when these verbs can participate in rare alternations. Perkins et al. (2022) had categorized these verbs as 'true' alternators following Levin 1993.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

Verb	Total	% Direct objects
fit	227	22%
play	1,568	19%
wait	383	15%
stand	294	7%
Intransitive		
run	228	6%
walk	253	4%
jump	197	4%
swim	180	4%
work	256	4%
cry	275	3%
sleep	451	3%
sit	859	1%
stay	308	1%
fall	605	0%

Table 2. *Known verbs and transitivity categories assumed by learner*

Type	Features
Object	Direct object of known verb is overt in canonical object position (right NP sister of V)
Subject	Subject of known verb is overt in canonical subject position (left NP sister of VP); sentence-initial; preceded by an auxiliary; preceded by another noun
Verb	Known verb is first verb in sentence; followed by a preposition or particle; has *-ed*, *-en*, *-ing*, *-s*, or irregular morphology
Tense & auxiliaries	Verb is preceded by *to*, *be*, *have*, *get*, or occurs with *do*
Other	Question; unknown function word sentence-initially, sentence-medially before verb, sentence-medially after verb, or sentence-finally

Table 3. *Direct objects and morphosyntactic features observed by learner (*X *and* F*). The presence of a direct object is the sole feature encoded by* X*. The remaining twenty-one features are encoded within the feature vector* $\vec{F}$*.*

learner's inference is driven by information relevant to the predicate-argument structure of a sentence: morphosyntactic features pertaining to subjects, objects, and verbs. These features are listed in Table 3.

In selecting these features, we model a learner with the representational abilities of an infant between the ages of fifteen and eighteen months. Prior behavioral evidence finds that infants at these ages can use the word-order properties of their language to identify clause subjects and objects in their canonical positions (Gagliardi et al. 2016, Hirsh-Pasek & Golinkoff 1996, Jin & Fisher 2014, Lidz et al. 2017, Perkins & Lidz 2020, Seidl et al. 2003). They attend to auxiliaries and can detect when the order of a subject and auxiliary is inverted (Geffen & Mintz 2015). They are able to segment a variety of verbal suffixes in English and other languages (Figueroa & Gerken 2019, Höhle et al. 2006, Kim & Sundara 2021, Mintz 2013, Nazzi et al. 2011, Santelmann & Jusczyk 1998, Soderstrom et al. 2002, Soderstrom et al. 2007, van Heugten & Shi 2010). In addition to auxiliaries and verbal affixes, infants at these ages are sensitive to the syntactic properties of a handful of other functional categories: determiners (Cauvet et al. 2014, Hicks et al. 2007, Höhle et al. 2004, Shi & Melançon 2010), pronouns (Cauvet et al. 2014), prepositions (Lidz et al. 2017), and negators (de Carvalho et al. 2021). Although they may not know the categories of other functional elements, they are able to recognize them as functional as opposed to lexical on the basis of their phonetic and prosodic properties (Monaghan et al. 2005, Shi et al. 1998, Shi et al. 1999).

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

In coding for the features in Table 3, we model an infant who can identify objects locally after verbs, but cannot yet identify nonlocal objects, such as fronted WH-phrases in WH-questions (Perkins & Lidz 2021). This means that sentences like *You're eating* and *What are you eating?* were both coded as not having a direct object from our learner's perspective, even though the WH-word *what* acts as a nonlocal object in the second sentence of this pair. Instead, WH-words are coded as 'unknown function words', a hypercategory that includes all functional elements assumed to be unknown at this age: WH-words, complementizers, quantifiers, focus particles, and conjunctions other than *and.*

We also code for the pragmatic feature 'question', which represents whether an utterance has interrogative force. Empirical evidence suggests that infants in their second year of life understand when a speaker is seeking information (Casillas & Frank 2017, Goodhue et al. 2023, Luchkina et al. 2018); see Carruthers 2018 on 'questioning attitudes' as a basic component of human minds. They do so likely on the basis of distributional, prosodic, and sociopragmatic cues (such as pauses and eye gaze) that differentiate questions from assertions in child-directed speech (Yang 2022). Young infants are sensitive to the prosodic and distributional differences between declaratives and polar questions (Frota et al. 2014, Geffen & Mintz 2015, Soderstrom et al. 2011). Although WH-questions differ from polar questions in their prosody (Geffen & Mintz 2017), it is possible that infants may know that these sentences are interrogatives, even before they are aware that they contain WH-dependencies (Gagliardi et al. 2016, Perkins & Lidz 2020, Seidl et al. 2003). Questions were identified by the presence of a question mark in the transcription; this does not distinguish constituent questions from polar questions.

In coding for the feature 'question', we abstract away from the specific prosodic features that learners might rely on to distinguish interrogatives from declaratives, and WH-questions from polar interrogatives (Frota et al. 2014, Geffen & Mintz 2015, Gryllia et al. 2020, Soderstrom et al. 2011, Yang 2022), which were not available in the corpora of child-directed speech used for our model's data set. In abstracting away from the prosodic signal, we ask how far a learner might get on the basis of distributional morphosyntactic information. However, we do not intend this as a claim that children cannot or do not additionally attend to this richer prosodic information, and further work might extend the current model to operate over a prosodically enriched data set.

To verify the accuracy of our automated coding, a random sample of 500 sentences from the data set were separately hand-coded by two trained researchers. Percentage agreement between the hand-coding and automated coding ranged from 87–100% across the twenty-one features; interrater reliability was also 87–100%. See Appendix B for more detail.

The sentences in the data set were also coded for their underlying clause types,[6] listed in Table 4. These annotations were used as a gold standard to evaluate our model and were not part of the model's data set. These clause types included three with movement: WH-questions, passives, and relative clauses. A given clause might be coded as multiple types: for example, as both a question and a passive. For sentences with multiple clauses, coding was conducted for the clause containing the verb of interest. Accuracy of clause-type coding was again evaluated by comparing against a 500-sentence sample hand-coded by two researchers. Percentage agreement between the hand-coding and automated coding ranged from 84–99% across the nine clause types (interrater reliability 87–99%); see Appendix B. Additional hand-coding was conducted for WH-questions and relative clauses in order to annotate the gap site in these sentences, which could not be reliably identified automatically for the entire data set.

4.2. Results

4.2.1. Sentence category distributions

Our joint inference model inferred thirty-nine total sentence categories, sixteen with transitivity violations and twenty-three without. To determine which of the model's inferred transitivity-violating categories

[6] We use 'clause type' as a theory-neutral term to refer to the various syntactic constructions in the clauses comprising our model's data set, in a way that is not directly related to the literature on the relation between clause types and speech acts (e.g. König & Siemund 2007, Sadock & Zwicky 1985).

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

Clause type	# Clauses	Description
Basic transitive	2,855 (15%)	Matrix, finite, declarative clause with overt direct object following known verb
Basic intransitive	2,704 (15%)	Matrix, finite, declarative clause without overt direct object following known verb
WH-question	2,336 (13%)	Clause has canonical syntactic form of a WH-question, with WH-element in a dependency with the known verb
Polar question	3,641 (20%)	Clause has canonical syntactic form of a polar question
Other question	1,922 (10%)	Clause was transcribed with a question mark, but does not have canonical syntactic form of a WH-question or polar question: includes tag, fragment, and echo questions, and rising-intonation declaratives
Passive	268 (1%)	Known verb has been passivized, excluding forms that are clearly adjectival
Relative clause	298 (2%)	Known verb is in a full or reduced relative clause
Other embedded clause	4,905 (27%)	Known verb is in a finite or nonfinite embedded, nonrelative clause.
Imperative	2,176 (12%)	Clause has canonical syntactic form of an imperative.

Table 4. *Distribution of underlying clause types in data set.*

were ones that contained object gaps (versus other types of transitivity violations), we calculated the odds ratio of direct objects appearing in these categories. This measure divides the odds of observing a feature in a given category by the odds of observing that feature outside of that category; an odds ratio significantly greater than 1 indicates that a feature is more likely to be present within than outside of the category, and an odds ratio significantly less than 1 indicates that a feature is more likely to be absent. Significance was calculated using a Fisher's exact test with a Bonferroni correction for multiple comparisons. See Appendix C for full details.

Of the sixteen transitivity-violating categories, fifteen had significantly lower odds (odds ratio less than 1) of producing direct objects; we call these 'object gap' categories. For each of the model's categories, Figure 2 displays the proportion of the category made up of each underlying clause type. Note that these proportions do not necessarily sum to 1 because a single clause might be of multiple types. For example, the sentences in the model's category 1 are entirely (1.00) WH-questions; this means that a given sentence in category 1 has a 100% probability of being tagged with the WH-question type in the gold-standard annotation. However, in category 2, a given sentence has a 95% probability of being a WH-question and also a 99% probability of being an embedded clause: this is a category that is predominantly long WH-questions, that is, those with WH-dependencies into embedded clauses.

In order to see whether the sentences in a given category predominantly belong to a particular clause type, versus being spread out among many different clause types, we calculated the purity of these categories when compared to the true underlying clause types in the corpora. Purity was calculated by counting the total number of sentences that belong to the predominant clause type in each category and dividing by the total number of sentences in the data set (Manning et al. 2008). Because a given sentence could belong to more than one clause type (i.e. both a WH-question and an embedded clause), we counted it as belonging to the predominant type in the category if that type was among those that the sentence belongs to. We note that this is a coarse approach and intend it only as a descriptive measure; our goal is not to evaluate the model on its clustering, but rather to evaluate it on whether it is able to find movement in its data, which we report in the following section. Given this approach, this measure has a minimum value of 0 if clusters are made up of a mixture of clause types, and a maximum value of 1 if clusters are made up of a single clause type. Our model's overall cluster purity is 0.76, which tells us that the model's categories were more likely to track one underlying clause type rather than a mixture.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

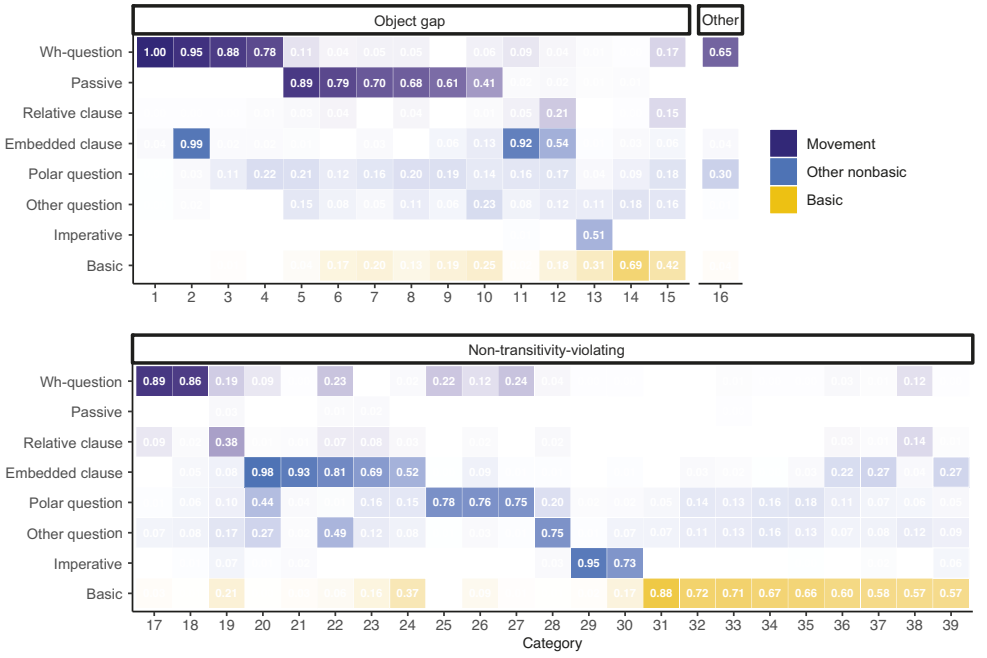

Figure 2. *Proportions of clause types in inferred sentence categories, joint inference model.*

The model inferred many more categories than necessary to identify the set of underlying clause types that it is being evaluated against. This is unsurprising: the learner was not given any information about how many clause-type categories were present, nor the grain size at which to perform its categorization. Instead, it was given leeway to posit as many categories as needed to explain the distributions of features and transitivity violations in its data. The model divided WH-questions among seven different categories: five transitivity-violating categories and two with no transitivity violations. These categories differentiate monoclausal from biclausal questions (e.g. *What does he eat?* vs. *What would you like to read?*), questions in the progressive aspect (e.g. *What are you bringing?*) from those in other aspects, and questions where the WH-word is sentence-initial from those where it is not (e.g. *And what is he wearing?*). The model also categorized subject questions separately from object and adjunct questions, and correctly identified subject questions as non-transitivity-violating. These distinctions may have implications for the learner's ability to generalize about the surface forms that are distinctive of different types of movement dependencies, a point we return to in the following sections.

4.2.2. Accuracy on identifying object movement

Here, we ask how well our learner can identify instances of object movement in its data. Visually, we can see from Figure 2 that clause types with movement were more likely to be categorized in object-gap categories than in non-object-gap categories. To ask how well the model identified cases of object movement specifically, we compared its object-gap categories against the sentences that were coded as actually having object gaps in the corpus. The model's accuracy is displayed in Figure 3 using three metrics. Precision measures the proportion of sentences in the model's object-gap categories that contained object movement according to our gold standard—that is, the proportion of these categories made up of object WH-questions, object relative clauses, or passives. Recall measures the proportion of sentences with object movement in the corpus overall that were identified as belonging to one of the model's object-gap categories. These metrics are not always aligned: it would be possible to achieve perfect recall by identifying all sentences as having object movement, but this would result in very poor precision. The F1 score, the harmonic mean of precision and recall, reflects the model's overall accuracy

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

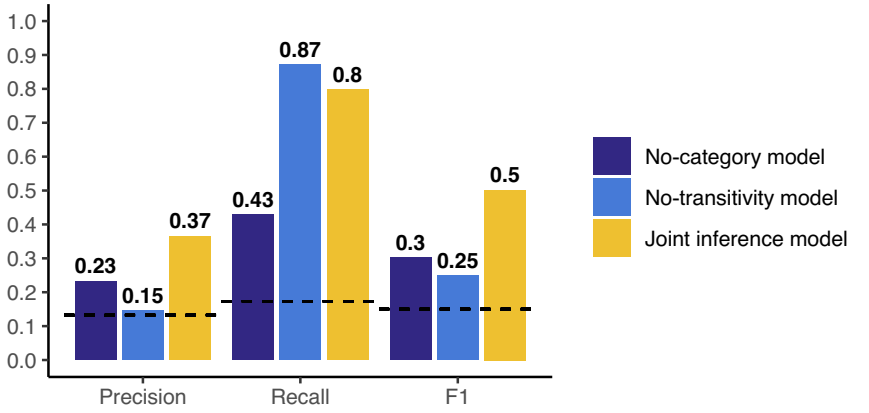

Figure 3. *Accuracy on identifying sentences with object movement in three metrics: precision (proportion of model's object-gap categories that contain object movement), recall (proportion of object movement in corpus identified by model), and F1 (harmonic mean of precision and recall).*

by taking into account both of these metrics. For each of these metrics, we compare the model's performance to a chance baseline, indicated by the dashed horizontal line. This represents the expected performance of a learner that randomly categorizes sentences as having transitivity violations that cause direct object gaps, by flipping a coin with weight 0.19, which is the probability of transitivity violations encoded in our learner's prior.

The model achieved an F1 score of 0.50. Its recall was 0.80, indicating that it identified 80% of sentences with object movement in its data. This accuracy rate is substantially above chance performance. Its precision was 0.37, indicating that on average, 37% of the sentences within its object-gap categories had instances of object movement. This precision rate is also above chance, but shows us that the model did not always manage to isolate object movement from other clause types in its data. To examine this further, we plotted the distribution of movement and nonmovement types in the model's object-gap categories in Figure 4. Object movement was the predominant clause type in 60% of these categories, but occurred alongside other movement types as well, particularly adjunct movement. The model appears to categorize adjunct movement together with object movement based on some surface distributional similarities: unlike subject movement, both object and adjunct movement contain subject-auxiliary inversion and can trigger *do*-support, even though adjunct movement does not tend to produce transitivity violations. The other 40% of the model's object-gap categories predominantly comprised sentences without movement. Thus, while the learner achieved high accuracy in identifying sentences with object movement as such, in certain cases it categorized sentences with object movement together with other clause types.

The model achieves this performance despite several factors that limit its accuracy. First, the model does not receive credit for identifying cases of movement other than WH-questions, passives, and relative clauses; other rarer cases of movement were more difficult to code automatically, and thus were not annotated in the gold-standard labels.[7] Second, the model infers object movement only from sentences that it believes violate verb transitivity: sentences with missing direct objects for verbs that it considers fully transitive. This means that the current evaluation measures how well the model was able to generalize from fully transitive verbs to verbs that also allow intransitive uses. Table 5 displays the proportions of sentences with object movement that the model correctly identified as having object gaps, broken down by the verb classes that comprised the model's prior transitivity knowledge. The model achieved high recall even though the majority of sentences with object movement occurred with verbs that it believed to be alternating, rather than obligatorily transitive. Of the 1,369 sentences coded as having

[7] These rarer movement types included *tough*-movement, movement out of purposive clauses, clefting, pseudo-clefting, topicalization, and comparative movement.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

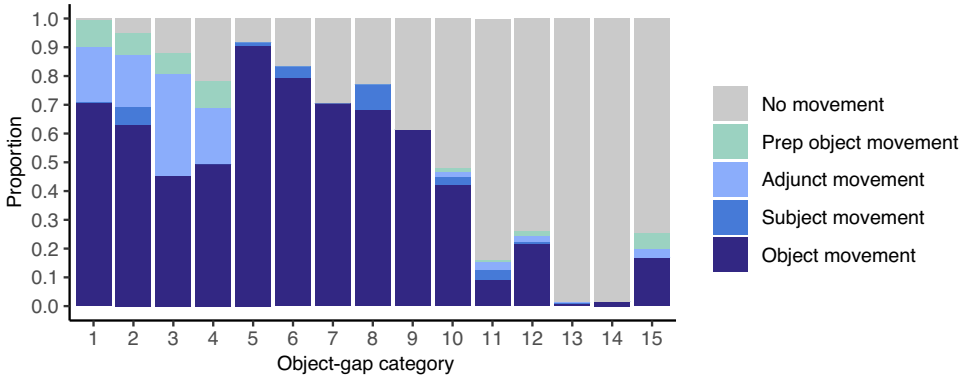

Figure 4. *Distribution of movement types in model's object-gap categories.*

Verb class	# Object-movement sentences	% Identified
Transitives	299	0.75
Intransitives	15	0.60
Alternators	1,055	0.82
Total	*1,369*	*0.80*

Table 5. *Proportion of object-movement sentences identified, by verb type.*

object movement in the corpus, only 299 contained known transitive verbs, compared to 1,055 containing known alternating verbs.[8] Nonetheless, the model achieved high accuracy across both the transitive and alternating verb classes. This tells us that it was able to generalize effectively: it used the presence of object gaps with known transitive verbs to identify the forms that object movement takes in its data, even with verbs that do not obligatorily require objects.[9]

In summary, our joint inference model performed significantly higher than chance in categorizing sentences with object movement in its data. It achieved a high recall rate, indicating that it was correctly able to identify the large majority of sentences with object movement that it encountered. Its accuracy was high for both transitive and alternating verbs, indicating that it was able to use the presence of transitivity violations with fully transitive verbs to identify direct object gaps with verbs that do not require objects. However, this object-gap inference produced a mixture of signal and noise: the sentences that the model categorized together with object movement also contained a variety of other movement and nonmovement clause types. This has potential implications for how informative the learner's categories are for isolating object movement from other syntactic dependencies, a question we turn to next.

4.2.3. Identifying distinctive features of object movement

Under our hypothesis, the sentence categories inferred by the joint inference model are an intermediate step of learning. Jointly inferring how to categorize sentences according to their surface features, and

[8] The few cases of object movement with intransitive verbs were uses of the verb in a rare or ungrammatical transitive frame (e.g. *What did you run?*).

[9] We note that each of the model's categories contains a mixture of different verbs (median number of verbs per category: 41, range: 10–50). Each category also includes a mixture of transitive, intransitive, and alternating verbs, in proportions similar to the overall proportions of these verb types in the data set (median proportions in model's categories: 0.21 transitive, 0.19 intransitive, 0.62 alternating; overall proportions in data set: 0.20 transitive, 0.20 intransitive, 0.61 alternating). The category that most closely tracks one verb and one verb type is category 8, which is predominantly (0.96) alternating verbs, and predominantly (0.77) the verb *break.* This is a category of passives that most frequently have the verb *broken.*

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

which sentence categories contain object gaps, helps a learner identify the particular forms that characterize different types of object movement in the target language. Here, we ask how well the model identified which specific surface features are the footprints of object movement. To do this, we assessed which surface features are most distinctive in the categories that the model inferred to have object gaps. If these include the characteristic forms of English object movement dependencies, then the model's sentence categories contain helpful information for identifying the ways that object movement can be realized in English.

To assess feature distinctiveness, we again calculated the odds ratio of each surface feature in the model's argument-gap categories. Table 6 reports the features with odds ratios significantly greater than 1 for each of the model's object-gap categories; full details are provided in Appendix C. Among these features are the characteristic forms of object movement dependencies in English. The categories that are predominantly WH-questions have greater odds of including subject-auxiliary inversion, *do*, and unknown function words sentence-initially or sentence-medially before the verb: these are WH-words. The categories predominantly made of passives have greater odds of including *get* or *be*, and *-en*, *-ed*, or irregular verbal morphology.

However, the distinctive features of object-gap categories also include forms that are irrelevant to movement dependencies. These include many positional characteristics of subjects and verbs, but also some specific morphemes. For instance, *be* and *-ing* are distinctive of two of the model's WH-question

Category	Primary clause type	Distinctive features
1	WH-question	Subject is overt, preceded by an aux; verb is first in sentence, has *-ing*, preceded by *be*; sentence-initial function word; question
2	WH-question & embedded	Verb is preceded by *to*; sentence-initial function word; question
3	WH-question	Subject is overt, preceded by an aux; verb is first in sentence, occurs with *do*; sentence-initial function word; question
4	WH-question	Subject is overt, preceded by an aux; verb is first in sentence, has *-ing*, preceded by *be*; sentence-medial function word before verb; question
5	passive	Verb has *-ed*, *-en*, or irregular form, preceded by *get*
6	passive	Verb has *-ed* or irregular form, preceded by *to* and *be*
7	passive	Subject (when overt) is sentence-initial; verb is first in sentence, has irregular form, preceded by *be* or *have*
8	passive	Subject is overt, sentence-initial; verb is first in sentence, verb has *-en* form, preceded by *be* or *have*
9	passive	Subject is overt, preceded by an NP; verb has *-en* or irregular form, preceded by *be* or *have*
10	passive	Subject (when overt) is sentence-initial; verb is first in sentence, has *-ed* form, preceded by *be* or *have*
11	embedded	Verb preceded by *to*; sentence-medial function word before verb
12	embedded	Subject is overt, preceded by an NP; verb has *-ed*, *-s*, or irregular form; sentence-medial function word before or after verb
13	imperative	Verb is first in sentence
14	basic	Subject is overt, sentence-initial; verb is first in sentence, has *-ed*, *-s*, or irregular form; function word sentence-finally or sentence-medially after verb
15	basic	Subject is overt, preceded by an NP; verb has *-ing*, preceded by *be*; sentence-medial function word before verb

Table 6. *Features with significantly higher odds in object-gap categories.*

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

categories, and *have* is distinctive of several of the model's passive categories. These features mark the realization of aspectual dependencies: *be* and *-ing* mark the progressive aspect, and *have* together with *-ed* or *-en* marks the perfect aspect. Thus, the model's categories contain both signal and noise for learning which surface features are the footprints of movement rather than other syntactic dependencies.

In summary, the current learner successfully identified the forms that characterize the most frequent types of movement in English, but it also identified some irrelevant features that are accidentally correlated with these forms. This invites the question of how a learner could effectively use this information for further steps of learning—how a learner could separate signal from noise by explaining some correlations as movement and others as different dependencies. It is possible that the model's ability to posit a potentially unbounded number of categories pushed it toward categories that are overly specific. Future work might test hypotheses about limits on the number of categories that a learner can posit, relaxing the ideal-learner assumption of this model in favor of one that more closely reflects the cognitive constraints that a child is operating within. It is unknown, however, whether this would lead the learner away from the accidental correlations that it identifies when its number of categories is unconstrained. Alternatively, it is also possible that a more sophisticated distributional learning mechanism might perform better. Further investigation is needed to determine whether the signal-to-noise ratio in the model's categories improves if it infers argument gaps using not only missing direct objects, but also other required but missing arguments (subjects and prepositional objects). This would give the learner the opportunity to identify nonobject movement; it is an open question whether this could make its inference about categories with argument gaps more precise.

4.3. Model comparisons

Our model achieves above-chance performance on identifying sentences with object movement by jointly inferring two properties: how sentences should be categorized together according to their surface feature distributions, and which sentence categories violate expectations about verb transitivity. To evaluate how important this joint inference is, we compare our model to baseline learners that perform only one step of inference at a time.

4.3.1. No-Category baseline

If it did not matter that our learner categorized sentences according to their surface features, then a learner should do just as well at identifying object movement on a sentence-by-sentence basis, by noting when objects are unexpectedly missing for known transitive verbs. To test whether the model's categorization process matters, we compared our model against a baseline learner that used only the presence or absence of direct objects in individual sentences, together with its knowledge of the transitivity properties of verbs in these sentences, to infer which sentences likely contain object gaps. Like our learner, this baseline model infers that an object gap is present when a transitive verb is unexpectedly missing its object. Unlike our learner, this model does not cluster sentences into categories according to their surface features, so it cannot draw inferences about which features are likely to be distinctive of sentences with object movement and cannot generalize the likely presence of an object gap from one sentence to another based on the similarity of their features.

This baseline learner has the architecture of the filtering model in Perkins et al. 2022, shown in Figure 5a. This is similar to the generative model in Figure 1, but omits the variables c, F, and $\delta^{(F)}$. With the variable c omitted, the learner does not assume that its direct object observations are partially governed by latent categories of sentences; with F and $\delta^{(F)}$ omitted, the learner does not observe or draw any inferences about the distributions of other surface features of sentences. Here, the variables e and $\delta^{(X)}$ are not indexed by sentence category: e represents whether an individual sentence (rather than a sentence category) contains a transitivity violation, and $\delta^{(X)}$ represents the rate of direct objects in individual sentences (rather than sentence categories) where transitivity violations are present. The learner's inference procedure consists of learning which sentences contain transitivity violations given its assumptions about verb transitivity and the rate of violations, but no joint learning about sentence categories on

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

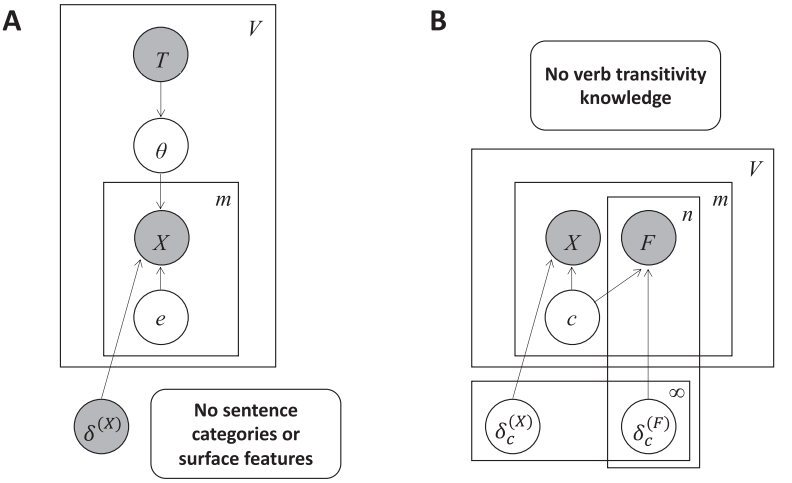

Figure 5. *Graphical models for (a) no-category baseline and (b) no-transitivity baseline.*

the basis of their distributional features. We fixed the transitivity properties T of each verb and the parameter $\delta^{(X)}$ to the values inferred by the learner in Perkins et al. 2022. We then sampled transitivity violations for each sentence in the corpus from the posterior probability distribution over e given X, T, and $\delta^{(X)}$, integrating over θ. See Appendix A for details.

To determine how well this 'no-category baseline' identified movement, we compared the sentences without direct objects that it inferred to have transitivity violations against the actual cases of object movement in the corpus. Its precision, recall, and F1 score are reported in Figure 3 above. The model achieved above-chance accuracy overall, but scored substantially lower than the joint inference model on all three metrics. This is because the baseline model's only source of reliable information for object gaps comes from the small percentage of verbs that it believes to be obligatorily transitive; it uses no other features in the sentences to inform this inference. If we examine its identification of object movement across verb classes, we find that it achieved high accuracy (74%) on identifying object movement with fully transitive verbs. But for the much larger percentage of verbs that are alternating, it can only guess which sentences contain gaps, identifying only 34% of object movement with these verbs. Thus, our joint inference model's ability to categorize sentences using a wide range of surface morphosyntactic features, and to generalize across sentences in a category, results in substantially better performance than inferring movement on a sentence-by-sentence basis from transitivity violations alone.

4.3.2. No-Transitivity baseline

Our second baseline comparison investigates how much prior verb transitivity knowledge constrains the learner's identification of movement—specifically, how important it is that our learner uses transitivity violations in the process of categorizing sentences by their surface morphosyntactic features. We compare our model against a learner that performs this categorization without knowing which verbs require direct objects. Like our learner, this baseline model uses the surface features of sentences to cluster sentences into categories. Unlike our learner, this baseline model does not have any knowledge of which verbs are transitive, so it cannot track transitivity violations in order to infer that object gaps are present in some of its sentence categories. Instead, it treats direct object observations identically to other surface features: for this learner, all direct objects are governed by the grammatical properties of a sentence category, not by the transitivity classes of verbs in the sentences. This learner therefore runs the risk of inferring categories that mix together sentences with movement and sentences without.

The architecture of this 'no-transitivity baseline' is shown in Figure 5b. This assumes the lowest portion of the generative model in Figure 1, omitting the variables T, θ, and e. When the variables T and θ are omitted, the learner now assumes that all direct object observations X are generated by $\delta_c^{(X)}$, the

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

grammatical properties of each sentence category, rather than by any properties of the verbs in these sentences. When the variable e is omitted, the learner no longer assumes that certain sentence categories contain transitivity violations. This means that its inference procedure consists of learning which sentence categories are present and which sentences belong to those categories, but no joint learning about transitivity violations in these categories. We sample category values for each sentence in the corpus from the posterior probability distribution over c given X and F, integrating over $\delta^{(X)}$ and $\delta^{(F)}$. See Appendix A for details.

Like our learner, the no-transitivity baseline inferred thirty-nine total categories. Of these, twenty-two had significantly lower odds of producing direct objects; we call these 'object-gap' categories, under the assumption that these are the learner's candidate categories for object movement. Full details are provided in Appendix C. The proportions of underlying clause types in the learner's categories are reported in Figure 6. These categories have similarly high purity to those inferred by the joint inference model: the baseline model's overall cluster purity is 0.77, compared to 0.76 for the joint inference model. This shows that the morphosyntactic features being tracked by both learners are informative for differentiating the various underlying clause types in the corpus, even without knowledge of which verbs require objects.

However, the baseline model's categories did not successfully differentiate sentences with movement from sentences without. The learner inferred many more sentence categories that were candidates for object movement, leading to slightly higher recall than our joint inference learner (Figure 3 above). But its precision was quite poor, leading to a substantially worse F1 score. To examine the source of this worse precision, we plotted the distribution of movement and nonmovement types in the model's object-gap categories in Figure 7. We find that object movement is the predominant clause type in only 27% of the learner's object-gap categories, compared to 60% in our joint inference learner. This tells us that our learner's ability to track transitivity violations is important for identifying categories of sentences with and without movement. While the distributions of morphosyntactic surface features of sentences convey a certain amount of information about the distinctions among different clause types, learning which of these

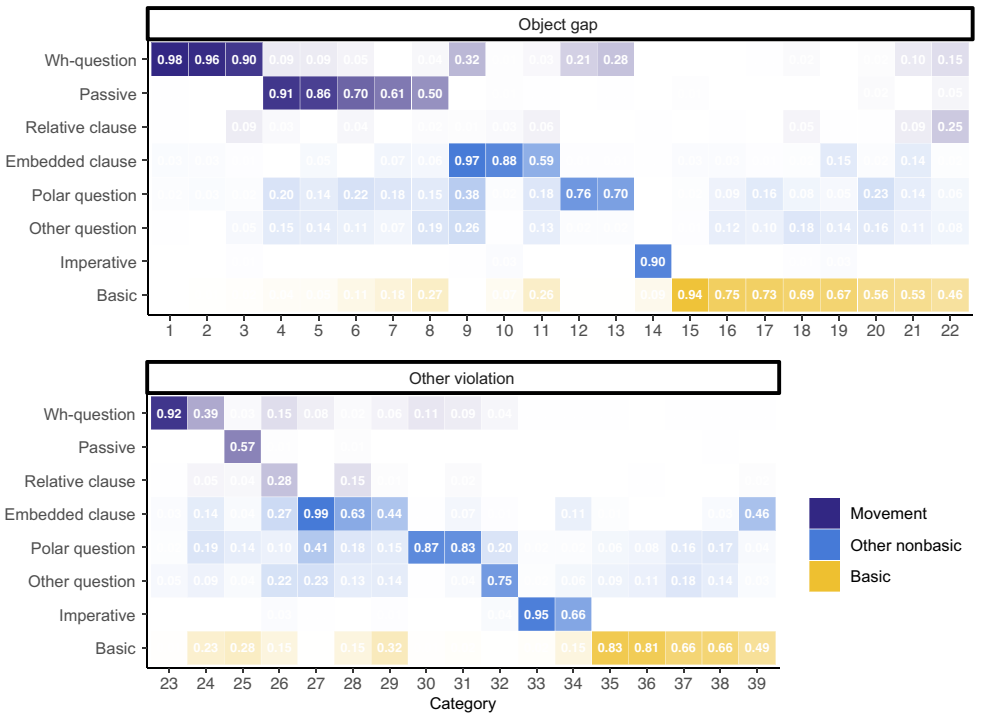

Figure 6. *Proportions of clause types in sentence categories, no-transitivity baseline.*

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

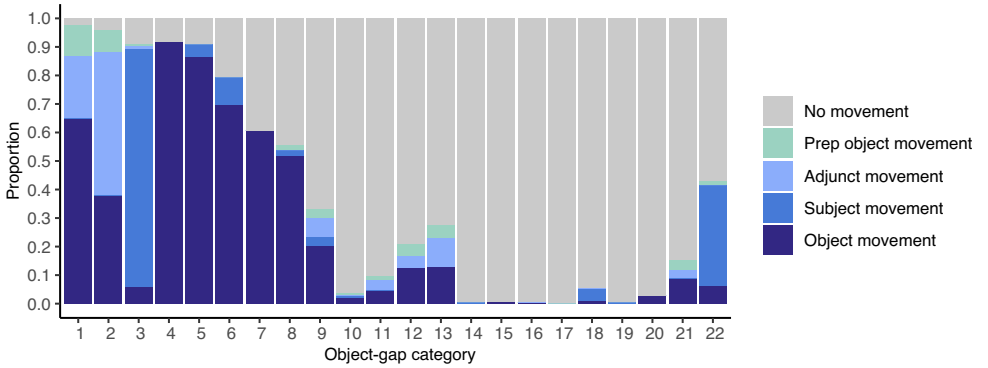

Figure 7. *Distribution of movement types in object-gap categories, no-transitivity baseline.*

distinctions signal movement, and which do not, requires the use of verb transitivity knowledge during distributional analysis.

4.4. Summary

In summary, our model identified 80% of sentences with object movement in child-directed speech, by tracking the surface morphosyntactic features of sentences that violate its expectations of verb transitivity. The model jointly infers how to categorize sentences according to their surface feature distributions and which of these sentence categories contain object gaps: unexpectedly missing objects of known verbs. This allowed the learner to generalize across sentences that share the same form and posit object gaps even for verbs that it does not know to be transitive. The learner performed substantially better than a baseline that relies only on known verb transitivity knowledge and does not categorize sentences on the basis of their surface feature distributions. This shows that the model's categorization process is important. It also outperformed a baseline that categorizes sentences using their surface features alone, without knowing which verbs require objects. The baseline learner performed substantially worse at differentiating sentences with and without object movement, showing that verb knowledge is an important guide for identifying movement.

5. General discussion

In order to acquire the system of syntactic dependencies in their language, children must detect evidence for abstract structure that is realized in highly variable ways within and across languages. Prior work has focused on how learners leverage statistical sensitivities to identify dependencies that are morphologically marked in their language (Gómez 2002, Gómez & Maye 2005, Höhle et al. 2006, Nazzi et al. 2011, Santelmann & Jusczyk 1998, Tincoff et al. 2000, van Heugten & Shi 2010). But these statistical learning mechanisms face challenges when encountering the fuller range of syntactic dependency types that learners must acquire. Movement dependencies provide an extreme example, both in their degree of abstraction and the degree of overt evidence available on the surface forms of sentences. How do learners identify a nonadjacent dependency between a fronted expression and the 'gap' of movement, which has no overt phonological form?

Here, we argue that solving this problem requires statistical learning not just over overt linguistic material, but also over hidden grammatical structure. Consistent with the literature on expectation violation in other domains of cognition (Denison & Xu 2012, Kouider et al. 2015, Stahl & Feigenson 2015, 2017, Téglás et al. 2011), we pursue the hypothesis that statistical learning is informed by unsatisfied grammatical predictions. When a learner encounters an unexpectedly missing predicted

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

argument of a verb, this may serve as evidence for a gap of an argument movement dependency. By tracking the surface forms that cooccur with these posited gap sites, learners may come to identify the distributional signatures of argument movement in the target language, enabling further inference about which specific syntactic dependencies underlie these surface forms. This hypothesis is motivated by prior empirical findings that knowledge of verb transitivity emerges before the identification of movement dependencies in infancy (Gagliardi et al. 2016, Jin & Fisher 2014, Lidz et al. 2017, Perkins & Lidz 2020, 2021).

Our findings demonstrate that this hypothesis is computationally feasible for the identification of object movement. Our learner jointly categorizes sentences according to similarities in their surface forms and infers which of these sentence categories violate its expectations about verb transitivity. This joint inference allows it to accurately identify the majority of object movement in child-directed speech and, in doing so, to identify the formal properties that are the footprints of object movement in English. It performs substantially better than baseline learners that rely on only one of these two sources of information: either learning from verb transitivity violations without using surface morphosyntactic features of sentences, or learning from distributions of surface features with no knowledge of verb transitivity. This shows that the learner's expectations about hidden grammatical structure, coming from prior verb argument-structure knowledge, place important constraints on its distributional learning mechanism. It thereby provides a computational account for why verb argument-structure knowledge developmentally precedes the acquisition of movement in a language like English.

These findings raise three sorts of questions for future research. First, how does a learner take information about the formal correlates of object gaps in the language and identify whether a particular form is realizing a movement dependency, versus another syntactic dependency? Our learner's inference yields both signal and noise for this next step of learning: the distinctive features of its object-gap categories include forms that characterize object movement in English, but also include forms that realize other nonmovement dependencies, such as aspectual dependencies. It is possible that children using this mechanism might be overly specific in the forms they associate with movement—for instance, inferring that progressive aspect is a hallmark of WH-questions, or perfect aspect is a hallmark of passives. Alternatively, perhaps a learner would identify fewer accidental correlations if the number of categories that it can posit for its data are limited, inviting further work exploring how children's developing cognitive capacities might interact with this type of distributional learning at young ages. But the current findings also raise the possibility that this learning mechanism is not sufficient to allow children to determine which expressions in a sentence are participating in movement dependencies, and which are not. Separating signal from noise may require supplementing information from formal distributions with additional information about the likely dependencies in a given sentence and the ways that those dependencies can be realized, so that a learner can successfully factor out the features that realize other dependencies from those that realize movement.

Prosody and pragmatics might provide additional relevant sources of information that are likely available to a young infant. Infants are sensitive to prosodic patterns from their first weeks of life (e.g. Christophe et al. 1994, Christophe et al. 2001, Gerken et al. 1994, Jusczyk et al. 1992, Nazzi et al. 1998). Because prosodic breaks tend to fall at the edges of syntactic phrases, past work has argued that infants may be able to use this information to help identify some of the constituency structure of an utterance (Christophe et al. 2008, de Carvalho et al. 2019, Gleitman et al. 1988, Gout et al. 2004, Morgan 1986, Morgan & Demuth 1996). Languages also deploy various other prosodic features, such as pitch and durational differences, to differentiate interrogatives from declaratives and WH-interrogatives from polar interrogatives (Frota et al. 2014, Geffen & Mintz 2015, Gryllia et al. 2020, Soderstrom et al. 2011, Yang 2022). Many of these features are language-specific and therefore must themselves be acquired, but it is possible that learners' inferences about the features that realize movement could be made more precise by tracking prosodic information in tandem with the morphosyntactic information provided to our model.

Infants also show early abilities to track the communicative intent of speakers (Csibra 2010, Meltzoff 1995, Woodward 2009) and to identify the speech act of an utterance, at least at a coarse level of granularity (Casillas & Frank 2017, Goodhue et al. 2023, Grosse et al. 2010, Liszkowski 2005, Luchkina et al. 2018). This speech-act information might also provide useful information about the syntactic

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

dependencies in a given sentence. However, as argued by Yang (2022), it is likely that this speech-act information would need to work in tandem with the type of syntactically guided distributional analysis proposed in the current work. Even a small amount of information about a speaker's communicative intent in using a particular sentence, along with the speaker's prosody, may help constrain the structure and interpretation that a learner assigns to that sentence. But it is likely that this information is not by itself constraining enough to provide a complete parse. Yang 2022 shows that it is computationally difficult to identify questions from child-directed speech using only pragmatic and prosodic information, and thus identifying which questions contain WH-dependencies would likely be even more challenging. This suggests that a learner might need to have available a partial syntactic representation for which this top-down information could be useful. This invites further investigation into how statistical learning might be supplemented both by a child's developing knowledge of possible syntactic dependencies and by knowledge of how those dependencies relate to speakers' goals in discourse.

A second important question for future research is how learners come to identify not only object WH-movement in their language, but also other types of movement dependencies. Our model uses only unexpectedly missing direct objects to infer when movement might be present, and therefore cannot identify subject, prepositional object, or adjunct movement. However, our model's exclusive focus on object movement is not intended as a claim that this form of movement must developmentally precede all others. Instead, these results merely demonstrate that the proposed expectation-violation learning mechanism could allow a learner to identify one type of argument movement that is empirically attested at early stages of development, while leaving open the possibility that other types of movement might also be acquired in tandem. In particular, it is possible that this mechanism could generalize to other forms of argument movement: in addition to tracking object gaps, a child might track when expected subjects or prepositional objects are unexpectedly absent, thereby allowing simultaneous inferences about the presence of subject and prepositional object movement. However, these gaps will be less obvious in matrix subject questions than in embedded questions, raising questions about the amount of evidence available for a learner to detect subject movement at young ages. A different learning mechanism would be required for the acquisition of adjunct movement, where no missing argument will signal the tail of the dependency.[10] Although some work finds that infants comprehend and produce subject and certain adjunct questions at young ages (Gagliardi et al. 2016, Perkins & Lidz 2020, Seidl et al. 2003, Stromswold 1995), further empirical work is needed to establish the developmental trajectory of infants' syntactic representations of these other forms of WH-movement relative to object movement, and to investigate the mechanisms by which they are acquired.

A third future research direction is determining how the proposed learning mechanism might generalize crosslinguistically. Our learner uses expectations about the word order of English to detect when direct objects are missing in their canonical positions. This hinges on the assumption that learners at this stage of development have already acquired some knowledge of how their language marks canonical predicate-argument relations. Some computational support exists for this assumption (Maitra & Perkins 2023, Perkins & Hunter 2023, 2026), but further empirical investigation is needed. In languages with a freer word order, other information, such as case morphology, may need to be recruited; see Fisher et al. 2019 and Suzuki & Kobayashi 2017 for evidence that Korean- and Japanese-learning two-year-olds are sensitive to this information in verb learning.

Moreover, using argument gaps as evidence for movement dependencies requires at least a reasonable correlation between empty arguments and movement in a language. This may be true for English (although see the caveat noted above for matrix subject questions), but this will be complicated in languages that allow syntactic null arguments or WH-in-situ. In languages like Korean and Japanese, learners must come to identify that many of the argument gaps they observe are null pronominals rather

[10] One possibility comes from the fact that our learner categorized adjunct questions together with object questions based on similarities in their surface morphosyntactic features: specifically, subject-auxiliary inversion and *do*-support. These features might serve as a cue that movement is present in adjunct questions, even though our learner currently identifies this as the wrong type of movement.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

than the gaps of movement; conversely, English learners must rule out a null pronominal analysis in favor of movement. And learners of WH-in-situ languages will not be able to rely on argument gaps in order to identify WH-dependencies; instead, they must come to recognize such dependencies even when the WH-element has not overtly moved to the clause position where it takes scope (Aoun et al. 1981, Huang 1982). It is possible that learners can more readily recognize when an in-situ WH-element bears a particular grammatical relation, but would need to use other formal, prosodic, or pragmatic information to recognize that this element is in a nonlocal dependency with a higher node in the clause, corresponding to the scope of the interrogative.

We suggest that the mechanism proposed here for English is one instance of a more general learning strategy that might be tailored to fit the evidence provided by a learner's data. Crosslinguistically, identifying canonical argument dependencies may be a necessary precursor to identifying nonlocal dependencies such as movement. An English learner may identify that word order provides a strong signal for canonical argument relations, and disruptions to this expected canonical word order signal that movement may be present. A Japanese learner may identify that case morphology is a better signal for these argument relations, that argument 'gaps' occur with a frequency that is more easily attributed to null pronominals rather than movement, and that overt and covert movement dependencies may be instead signaled by additional formal, prosodic, or pragmatic properties. In both cases, it is plausible that a learner's initial knowledge of the core predicate-argument structure of a clause provides an important grammatical scaffold for guiding future learning from the surface distributions in the data. This invites further empirical and computational work studying the developmental trajectory of argument structure and argument movement crosslinguistically.

More broadly, the current findings illustrate how two learning mechanisms with analogues in other areas of cognition—statistical learning and learning from expectation violation—can be combined to novel effect in the domain of language acquisition. On this proposal, prior grammatical knowledge creates expectations that, when violated, form the basis for inferring hidden grammatical structure. Statistical learning may then be conducted over this hidden structure as well as more observable forms in the data. Here, we suggest that this combination provides a powerful foothold into syntactic dependency learning in early language development. This may also provide new avenues for understanding how incremental learning proceeds in not only language acquisition but also other domains of cognition, where predictions generated from knowledge acquired earlier in development form part of the data that learners use to draw new generalizations.

Data availability statement. Code and data for the model and simulations reported in this article can be found at https://github.com/perkinsl/mind-the-gap/.

Acknowledgments. We thank Shounak Kuiry, Lillianna Righter, Jordan Schneider, and John-Paul Teti for assistance in coding and data preparation. We also thank Lisa Pearl, Alexander Williams, and audiences at BUCLD 2019 and the University of Maryland CNL Lab for their helpful feedback on earlier versions of this work. [Full editorial history: Received 01 June 2024; revision invited 05 April 2025; revision received 31 July 2025; accepted pending revisions 05 October 2025; revision received 09 January 2026; accepted 12 January 2026.]

Funding disclosure statement. This work was supported by the National Science Foundation (#BCS-1551629, Doctoral Dissertation Improvement grant #BCS-1827709, and NRT award #DGE-1449815), by the Division of Behavioral and Cognitive Sciences (#1551629, #1827709), and by the Division of Graduate Education (#1449815).

Conflict of interest. The authors declare no conflict of interest.

Ethics statement. None.

References

Abend, Omri; Tom Kwiatkowski; Nathaniel J. Smith; Sharon Goldwater; and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition* 164.116–43. https://doi.org/10.1016/j.cognition.2017.02.009.

Alishahi, Afra, and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science* 32(5).789–834. https://doi.org/10.1080/03640210801929287.

Anderson, John R. 1990. *The adaptive character of thought*. New York: Psychology Press. https://doi.org/10.4324/9780203771730.
Anderson, John R., and Michael Matessa. 1990. A rational analysis of categorization. *Machine Learning Proceedings 1990*, 76–84. https://doi.org/10.1016/B978-1-55860-141-3.50013-4.
Aoshima, Sachiko; Colin Phillips; and Amy Weinberg. 2004. Processing filler-gap dependencies in a head-final language. *Journal of Memory and Language* 51(1).23–54. https://doi.org/10.1016/j.jml.2004.03.001.
Aoun, Joseph; Norbert Hornstein; and Dominique Sportiche. 1981. Some aspects of wide scope quantification. *Journal of Linguistic Research* 1(3).69–95.
Berwick, Robert C. 1985. *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1074.001.0001.
Brown, Roger. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674732469.
Carruthers, Peter. 2018. Basic questions. *Mind & Language* 33(2).130–47. https://doi.org/10.1111/mila.12167.
Casillas, Marisa, and Michael C. Frank. 2017. The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language* 92.234–53. https://doi.org/10.1016/j.jml.2016.06.013.
Cauvet, Elodie; Rita Limissuri; Séverine Millotte; Katrin Skoruppa; Dominique Cabrol; and Anne Christophe. 2014. Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development* 10(1).1–18. https://doi.org/10.1080/15475441.2012.757970.
Cheng, Lisa Lai-Shen. 2003. Wh-in-situ. *Glot International* 7(4).103–9.
Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, Noam. 1980. On cognitive structures and their development: A reply to Piaget. *Language and learning: The debate between Jean Piaget and Noam Chomsky*, ed. by Massimo Piatelli-Palmarini, 35–54. Cambridge, MA: Harvard University Press.
Christophe, Anne; Emmanuel Dupoux; Josiane Bertoncini; and Jacques Mehler. 1994. Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *The Journal of the Acoustical Society of America* 95(3).1570–80. https://doi.org/10.1121/1.408544.
Christophe, Anne; Jacques Mehler; and Núria Sebastián-Gallés. 2001. Perception of prosodic boundary correlates by newborn infants. *Infancy* 2(3).385–94. https://doi.org/10.1207/S15327078IN0203_6.
Christophe, Anne; Séverine Millotte; Savita Bernal; and Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. *Language and Speech* 51(1–2).61–75. https://doi.org/10.1177/00238309080510010501.
Cole, Peter, and Gabriella Hermon. 1994. Is there LF wh-movement? *Linguistic Inquiry* 25(2).239–62. https://www.jstor.org/stable/4178860.
Crain, Stephen, and Janet Dean Fodor. 1985. How can grammars help parsers? *Natural language parsing: Psychological, computational, and theoretical perspectives*, ed. by David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky, 94–128. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.004.
Csibra, Gergely. 2010. Recognizing communicative intentions in infancy. *Mind & Language* 25(2).141–68. https://doi.org/10.1111/j.1468-0017.2009.01384.x.
Dayley, Jon P. 1981. *Tzutujil grammar*. Berkeley: University of California, Berkeley dissertation. http://escholarship.org/uc/item/4cx9t9b1.
de Carvalho, Alex; Cécile Crimon; Axel Barrault; John Trueswell; and Anne Christophe. 2021. 'Look! It is not a bamoule!': 18- and 24-month-olds can use negative sentences to constrain their interpretation of novel word meanings. *Developmental Science* 24(4):e13085. https://doi.org/10.1111/desc.13085.
de Carvalho, Alex; Angela Xiaoxue He; Jeffrey Lidz; and Anne Christophe. 2019. Prosody and function words cue the acquisition of word meanings in 18-month-old infants. *Psychological Science* 30(3).319–32. https://doi.org/10.1177/0956797618814131.
Denison, Stephanie, and Fei Xu. 2012. Probabilistic inference in human infants. *Advances in child development and behavior, vol. 43: Rational constructivism in cognitive development*, ed. by Tamar Kushnir, Janette B. Benson, and Fei Xu, 27–58. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-12-397919-3.00002-2.
Dillon, Brian; Ewan Dunbar; and William Idsardi. 2013. A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science* 37(2).344–77. https://doi.org/10.1111/cogs.12008.
Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14(2).179–211. https://doi.org/10.1207/s15516709cog1402_1.
Feldman, Naomi H.; Sharon Goldwater; Emmanuel Dupoux; and Thomas Schatz. 2021. Do infants really learn phonetic categories? *Open Mind* 5.113–31. https://doi.org/10.1162/opmi_a_00046.
Feldman, Naomi H.; Thomas L. Griffiths; Sharon Goldwater; and James L. Morgan. 2013. A role for the developing lexicon in phonetic category acquisition. *Psychological Review* 120(4).751–78. https://doi.org/10.1037/a0034245.
Ferguson, Thomas S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1(2).209–30. http://www.jstor.org/stable/2958008.
Figueroa, Megan, and LouAnn Gerken. 2019. Experience with morphosyntactic paradigms allows toddlers to tacitly anticipate overregularized verb forms months before they produce them. *Cognition* 191:103977. https://doi.org/10.1016/j.cognition.2019.05.014.
Fisher, Cynthia; Kyong-sun Jin; and Rose M. Scott. 2019. The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science* 12(1).48–77. https://doi.org/10.1111/tops.12447.
Fodor, Janet Dean 1998. Parsing to learn. *Journal of Psycholinguistic Research* 27(3).339–74. https://doi.org/10.1023/A:1023255705029.

Frank, Michael C.; Noah D. Goodman; and Joshua B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20(5).578–85. https://doi.org/10.1111/j.1467-9280.2009.02335.x.

Frazier, Lyn, and Charles Clifton, Jr. 1989. Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes* 4(2).93–126. https://doi.org/10.1080/01690968908406359.

Frazier, Lyn, and Giovanni B. Flores d'Arcais. 1989. Filler driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language* 28(3).331–44. https://doi.org/10.1016/0749-596X(89)90037-5.

Frota, Sonia; Joseph Butler; and Marina Vigário. 2014. Infants' perception of intonation: Is it a statement or a question? *Infancy* 19(2).194–213. https://doi.org/10.1111/infa.12037.

Gagliardi, Annie; Tara M. Mease; and Jeffrey Lidz. 2016. Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15- and 20-month-olds. *Language Acquisition* 23(3).234–60. https://doi.org/10.1080/10489223.2015.1115048.

Geffen, Susan, and Toben H. Mintz. 2015. Can you believe it? 12-month-olds use word order to distinguish between declaratives and polar interrogatives. *Language Learning and Development* 11(3).270–84. https://doi.org/10.1080/15475441.2014.951595.

Geffen, Susan, and Toben H. Mintz. 2017. Prosodic differences between declaratives and interrogatives in infant-directed speech. *Journal of Child Language* 44(4).968–94. https://doi.org/10.1017/S0305000916000349.

Geman, Stuart, and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6).721–41. https://doi.org/10.1109/TPAMI.1984.4767596.

Gerken, LouAnn; Peter W. Jusczyk; and Denise R. Mandel. 1994. When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition* 51(3).237–65. https://doi.org/10.1016/0010-0277(94)90055-8.

Gleitman, Lila R.; Henry Gleitman; Barbara Landau; and Eric Wanner. 1988. Where learning begins: Initial representations for language learning. *Linguistics: The Cambridge survey, vol. 3. Language: Psychological and biological aspects*, ed. by Frederick J. Newmeyer, 150–93. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621062.007.

Goldwater, Sharon; Thomas L. Griffiths; and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1).21–54. https://doi.org/10.1016/j.cognition.2009.03.008.

Gómez, Rebecca L. 2002. Variability and detection of invariant structure. *Psychological Science* 13(5).431–36. https://doi.org/10.1111/1467-9280.00476.

Gómez, Rebecca L., and Jessica Maye. 2005. The developmental trajectory of nonadjacent dependency learning. *Infancy* 7(2).183–206. https://doi.org/10.1207/s15327078in0702_4.

Goodhue, Daniel; Valentine Hacquard; and Jeffrey Lidz. 2023. 18-month-olds understand the links between declaratives and assertions, and interrogatives and questions. *Proceedings of the Boston University Conference on Language Development (BUCLD)* 47.331–41. https://www.lingref.com/bucld/47/BUCLD47-27.pdf.

Gout, Ariel; Anne Christophe; and James L. Morgan. 2004. Phonological phrase boundaries constrain lexical access II: Infant data. *Journal of Memory and Language* 51(4).548–67. https://doi.org/10.1016/j.jml.2004.07.002.

Griffiths, Thomas L.; Nick Chater; and Joshua B. Tenenbaum. 2024. *Bayesian models of cognition: Reverse engineering the mind*. Cambridge, MA: MIT Press.

Grosse, Gerlind; Tanya Behne; Malinda Carpenter; and Michael Tomasello. 2010. Infants communicate in order to be understood. *Developmental Psychology* 46(6).1710–22. https://doi.org/10.1037/a0020727.

Gryllia, Stella; Jenny S. Doetjes; Yang Yang; and Lisa Lai-Shen Cheng. 2020. Prosody, clause typing, and *wh*-in-situ: Evidence from Mandarin. *Laboratory Phonology* 11(1):19. https://doi.org/10.5334/labphon.169.

Hicks, Jessica; Jessica Maye; and Jeffrey Lidz. 2007. The role of function words in infants' syntactic categorization of novel words. Paper presented at the annual meeting of the Linguistic Society of America, Anaheim, CA.

Hirsh-Pasek, Kathy, and Roberta Michnick Golinkoff. 1996. The intermodal preferential looking paradigm: A window onto emerging language comprehension. *Methods for assessing children's syntax*, ed. by Dana McDaniel, Cecile McKee, and Helen Smith Cairns, 105–24. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/4575.003.0009.

Hirzel, Mina; Laurel Perkins; and Jeffrey Lidz. 2020. 19 month-olds represent and incrementally parse filler-gap dependencies. Paper presented at the 33rd Annual CUNY Human Sentence Processing Conference. https://osf.io/v3k27/.

Höhle, Barbara; Michaela Schmitz; Lynn M. Santelmann; and Jurgen Weissenborn. 2006. The recognition of discontinuous verbal dependencies by German 19-month-olds: Evidence for lexical and structural influences on children's early processing capacities. *Language Learning and Development* 2(4).277–300. https://doi.org/10.1207/s15473341lld0204_3.

Höhle, Barbara; Jürgen Weissenborn; Dorothea Kiefer; Antje Schulz; and Michaela Schmitz. 2004. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy* 5(3).341–53. https://doi.org/10.1207/s15327078in0503_5.

Huang, C.-T. James. 1982. *Logical relations in Chinese and the theory of grammar*. Cambridge, MA: MIT dissertation. http://hdl.handle.net/1721.1/15215.

Jin, Kyong-sun, and Cynthia Fisher. 2014. Early evidence for syntactic bootstrapping: 15-month-olds use sentence structure in verb learning. *Boston University Conference on Language Development (BUCLD) 38 Online Proceedings Supplement*. https://www.bu.edu/bucld/files/2014/04/jin.pdf.

Jusczyk, Peter W.; Kathy Hirsh-Pasek; Deborah G. Kemler Nelson; Lori J. Kennedy; Amanda Woodward; and Julie Piwoz. 1992. Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology* 24(2).252–93. https://doi.org/10.1016/0010-0285(92)90009-Q.

Kim, Yun Jung, and Megha Sundara. 2021. 6-month-olds are sensitive to English morphology. *Developmental Science* 24(4): e13089. https://doi.org/10.1111/desc.13089.

König, Ekkehard, and Peter Siemund. 2007. Speech act distinctions in grammar. *Language typology and syntactic description*, ed. by Timothy Shopen, 276–324. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511619427.005.

Kouider, Sid; Bria Long; Lorna Le Stanc; Sylvain Charron; Anne-Caroline Fievet; Leonardo S. Barbosa; and Sofie V. Gelskov. 2015. Neural dynamics of prediction and surprise in infants. *Nature Communications* 6(1):8537. https://doi.org/10.1038/ncomms9537.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Lidz, Jeffrey, and Annie Gagliardi. 2015. How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics* 1(1).333–53. https://doi.org/10.1146/annurev-linguist-030514-125236.

Lidz, Jeffrey; Aaron Steven White; and Rebecca Baier. 2017. The role of incremental parsing in syntactically conditioned word learning. *Cognitive Psychology* 97.62–78. https://doi.org/10.1016/j.cogpsych.2017.06.002.

Liszkowski, Ulf. 2005. Human twelve-month-olds point cooperatively to share interest with and helpfully provide information for a communicative partner. *Gesture* 5(1–2).135–54. https://doi.org/10.1075/gest.5.1.11lis.

Luchkina, Elena; David M. Sobel; and James L. Morgan. 2018. Eighteen-month-olds selectively generalize words from accurate speakers to novel contexts. *Developmental Science* 21(6):e12663. https://doi.org/10.1111/desc.12663.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk, vol. 2: The database*. New York: Psychology Press. https://doi.org/10.4324/9781315805641.

Maitra, Shalinee, and Laurel Perkins. 2023. Filtering input for learning constrained grammatical variability: The case of Spanish word order. *Proceedings of the Society for Computation in Linguistics 2023*, 108–20. https://aclanthology.org/2023.scil-1.9/.

Manning, Christopher; Prabhakar Raghavan; and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman and Co.

Maye, Jessica; Janet F. Werker; and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82(3).B101–B111. https://doi.org/10.1016/S0010-0277(01)00157-3.

McMurray, Bob; Richard N. Aslin; and Joseph C. Toscano. 2009. Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science* 12(3).369–78. https://doi.org/10.1111/j.1467-7687.2009.00822.x.

Meltzoff, Andrew N. 1995. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31(5).838–50. https://doi.org/10.1037/0012-1649.31.5.838.

Mintz, Toben H. 2013. The segmentation of sub-lexical morphemes in English-learning 15-month-olds. *Frontiers in Psychology* 4: 24. https://doi.org/10.3389/fpsyg.2013.00024.

Monaghan, Padraic; Nick Chater; and Morten H. Christiansen. 2005. The differential role of phonological and distributional cues in grammatical categorisation. *Cognition* 96(2).143–82. https://doi.org/10.1016/j.cognition.2004.09.001.

Morgan, James L. 1986. *From simple input to complex grammar*. Cambridge, MA: MIT Press.

Morgan, James L., and Katherine Demuth (eds.) 1996. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Nazzi, Thierry; Isabelle Barrière; Louise Goyet; Sarah Kresh; Géraldine Legendre. 2011. Tracking irregular morphophonological dependencies in natural language: Evidence from the acquisition of subject-verb agreement in French. *Cognition* 120(1).119–35. https://doi.org/10.1016/j.cognition.2011.03.004.

Nazzi, Thierry; Josiane Bertoncini; and Jacques Mehler. 1998. Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance* 24(3).756–66. https://doi.org/10.1037//0096-1523.24.3.756.

Pearl, Lisa, and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition* 20(1).23–68. https://doi.org/10.1080/10489223.2012.738742.

Pearl, Lisa, and Jon Sprouse. 2019. Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. *Language* 95(4).583–611. https://doi.org/10.1353/lan.2019.0067.

Perfors, Amy; Joshua B. Tenenbaum; and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118(3).306–38. https://doi.org/10.1016/j.cognition.2010.11.001.

Perfors, Amy; Joshua B. Tenenbaum; and Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language* 37(3).607–42. https://doi.org/10.1017/S0305000910000012.

Perkins, Laurel. 2019. *How grammars grow: Argument structure and the acquisition of non-basic syntax*. College Park: University of Maryland dissertation.

Perkins, Laurel; Naomi H. Feldman; and Jeffrey Lidz. 2022. The power of ignoring: Filtering input for argument structure acquisition. *Cognitive Science* 46(1):e13080. https://doi.org/10.1111/cogs.13080.

Perkins, Laurel, and Tim Hunter. 2023. Noise-tolerant learning as selection among deterministic grammatical hypotheses. *Proceedings of the Society for Computation in Linguistics 2023*, 186–98. https://aclanthology.org/2023.scil-1.16/.

Perkins, Laurel, and Tim Hunter. 2026. Modeling regularization in language acquisition as noise-tolerant grammar selection. *Cognition* 268:106352. https://doi.org/10.1016/j.cognition.2025.106352.

Perkins, Laurel, and Jeffrey Lidz. 2020. Filler-gap dependency comprehension at 15 months: The role of vocabulary. *Language Acquisition* 27(1).98–115. https://doi.org/10.1080/10489223.2019.1659274.

Perkins, Laurel, and Jeffrey Lidz. 2021. 18-month-old infants represent nonlocal syntactic dependencies. *Proceedings of the National Academy of Sciences* 118(41):e2026469118. https://doi.org/10.1073/pnas.2026469118.
Pinker, Steven. 1984. *Language learnability and language development*. Cambridge, MA: Harvard University Press.
Reinhart, Tanya. 1998. Wh-in-situ in the framework of the minimalist program. *Natural Language Semantics* 6(1).29–56. https://doi.org/10.1023/A:1008240014550.
Sadock, Jerrold M., and Arnold M. Zwicky. 1985. Speech act distinctions in syntax. *Language typology and syntactic description*, ed. by Timothy Shopen, 155–96. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511619427.005.
Sakas, William G., and Janet Dean Fodor. 2001. The structural triggers learner. *Language acquisition and learnability*, ed. by Stefano Bertolo, 172–233. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511554360.006.
Sakas, William G., and Janet Dean Fodor. 2012. Disambiguating syntactic triggers. *Language Acquisition* 19(2).83–143. https://doi.org/10.1080/10489223.2012.660553.
Sanborn, Adam N.; Thomas L. Griffiths; and Richard M. Shiffrin. 2010. Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology* 60(2).63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001.
Santelmann, Lynn M., and Peter W. Jusczyk. 1998. Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition* 69(2).105–34. https://doi.org/10.1016/S0010-0277(98)00060-2.
Seidl, Amanda; George Hollich; and Peter W. Jusczyk. 2003. Early understanding of subject and object wh-questions. *Infancy* 4(3).423–36. https://doi.org/10.1207/S15327078IN0403_06.
Shi, Rushen, and Andréane Melançon. 2010. Syntactic categorization in French-learning infants. *Infancy* 15(5).517–33. https://doi.org/10.1111/j.1532-7078.2009.00022.x.
Shi, Rushen; James L. Morgan; and Paul Allopenna. 1998. Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language* 25(1).169–201. https://doi.org/10.1017/S0305000997003395.
Shi, Rushen; Janet F. Werker; and James L. Morgan. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition* 72(2).B11–B21. https://doi.org/10.1016/S0010-0277(99)00047-5.
Soderstrom, Melanie; Megan Blossom; Rina Foygel; and James L. Morgan. 2008. Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language* 35(4).869–902. https://doi.org/10.1017/S0305000908008763.
Soderstrom, Melanie; Eon-Suk Ko; and Uliana Nevzorova. 2011. It's a question? Infants attend differently to yes/no questions and declaratives. *Infant Behavior and Development* 34(1).107–10. https://doi.org/10.1016/j.infbeh.2010.10.003.
Soderstrom, Melanie; Kenneth Wexler; and Peter W. Jusczyk. 2002. English-learning toddlers' sensitivity to agreement morphology in receptive grammar. *Proceedings of the Boston University Conference on Language Development (BUCLD)* 26.643–52.
Soderstrom, Melanie; Katherine S. White; Erin Conwell; and James L. Morgan. 2007. Receptive grammatical knowledge of familiar content words and inflection in 16-month-olds. *Infancy* 12(1).1–29. https://doi.org/10.1111/j.1532-7078.2007.tb00231.x.
Stahl, Aimee E., and Lisa Feigenson. 2015. Observing the unexpected enhances infants' learning and exploration. *Science* 348(6230).91–94. https://doi.org/10.1126/science.aaa3799.
Stahl, Aimee E., and Lisa Feigenson. 2017. Expectancy violations promote learning in young children. *Cognition* 163.1–14. https://doi.org/10.1016/j.cognition.2017.02.008.
Stromswold, Karin. 1995. The acquisition of subject and object wh-questions. *Language Acquisition* 4(1–2).5–48. https://doi.org/10.1080/10489223.1995.9671658.
Suppes, Patrick. 1974. The semantics of children's language. *American Psychologist* 29(2).103–14. https://doi.org/10.1037/h0036026.
Sussman, Rachel Shirley, and Julie Sedivy. 2003. The time-course of processing syntactic dependencies: Evidence from eye movements. *Language and Cognitive Processes* 18(2).143–63. https://doi.org/10.1080/01690960143000498.
Suzuki, Takaaki, and Tessei Kobayashi. 2017. Syntactic cues for inferences about causality in language acquisition: Evidence from an argument-drop language. *Language Learning and Development* 13(1).24–37. https://doi.org/10.1080/15475441.2016.1193019.
Téglás, Ernő; Edward Vul; Vittorio Girotto; Michel Gonzalez; Joshua B. Tenenbaum; and Luca L. Bonatti. 2011. Pure reasoning in 12-month-old infants as probabilistic inference. *Science* 332(6033).1054–59. https://doi.org/10.1126/science.1196404.
Tincoff, Ruth; Lynn M. Santelmann; and Peter W. Jusczyk. 2000. Auxiliary verb learning and 18-month-olds' acquisition of morphological relationships. *Proceedings of the Boston University Conference on Language Development (BUCLD)* 24.726–37.
Traxler, Matthew J., and Martin J. Pickering. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language* 35(3).454–75. https://doi.org/10.1006/jmla.1996.0025.
Valian, Virginia. 1990. Logical and psychological constraints on the acquisition of syntax. *Language processing and language acquisition*, ed. by Lyn Frazier and Jill G. de Villiers, 119–45. Dordrecht: Kluwer. https://doi.org/10.1007/978-94-011-3808-6_5.
Valian, Virginia. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition* 40(1–2).21–81. https://doi.org/10.1016/0010-0277(91)90046-7.
Vallabha, Gautam K.; James L. McClelland; Ferran Pons; Janet F. Werker; and Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences* 104(33).13273–78. https://doi.org/10.1073/pnas.0705369104.
van Heugten, Marieke, and Rushen Shi. 2010. Infants' sensitivity to non-adjacent dependencies across phonological phrase boundaries. *The Journal of the Acoustical Society of America* 128(5).EL223–EL228. https://doi.org/10.1121/1.3486197.

Wexler, Kenneth, and Peter Culicover. 1980. *Formal principles of language acquisition*. Cambridge, MA: MIT Press.
White, Aaron Steven, and Jeffrey Lidz. 2022. Lexicalization in the developing parser. *Glossa Psycholinguistics* 1(1). https://doi.org/10.5070/G601148.
Woodward, Amanda L. 2009. Infants' grasp of others' intentions. *Current Directions in Psychological Science* 18(1).53–57. https://doi.org/10.1111/j.1467-8721.2009.01605.x.
Yang, Charles D. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199254149.001.0001.
Yang, Yu'an. 2022. *Are you asking me or telling me? Learning clause types and speech acts in English and Mandarin*. College Park: University of Maryland dissertation. https://doi.org/10.13016/sfhx-z3h2.
Yuan, Sylvia; Cynthia Fisher; and Jesse Snedeker. 2012. Counting the nouns: Simple structural cues to verb meaning. *Child Development* 83(4).1382–99. https://doi.org/10.1111/j.1467-8624.2012.01783.x.

Appendix A: Details of Gibbs sampling

A1. Joint inference learner

We use Gibbs sampling (Geman & Geman 1984) to jointly infer c and e, integrating over θ, $\delta^{(X)}$, and $\delta^{(F)}$.

A1.1. Sampling c

To begin, values of c for each sentence are initialized to one of three initial sentence categories: one category with transitivity violations and two without. These initial categories are sampled from the posterior probability distribution that a given sentence contains a transitivity violation under the model in Perkins et al. 2022. This uses the same sampling equations as for the no-category baseline, reported in Section A2. If a sentence is sampled as containing a transitivity violation under that model, it is initialized to the transitivity-violating category; if not, it is randomly initialized to one of the two nonviolating categories. We used two nonviolating categories rather than one because this improved the sampler's convergence.

After initializing c, new values of c for each sentence are resampled sequentially. From observations of direct objects and other features in a sentence, and across other sentences in the model's data, the model determines which previously seen or new value of c was most likely to have generated those observations. For direct object observation $X_i^{(v)}$ and other feature observations $\vec{F}_i^{(v)}$ in sentence i, together with all other direct object observations $\mathbf{X}_{-i}$, feature observations $\vec{\mathbf{F}}_{-i}$, and sentence category assignments $\mathbf{c}_{-i}$ for other sentences in the data set, we use Bayes's rule to compute the posterior probability of each value for c.

$$p\left(c_i|X_i^{(v)},\vec{F}_i^{(v)},T^{(v)},e_c,\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)=\frac{p\left(X_i^{(v)},\vec{F}_i^{(v)}|c_i,e_c,T^{(v)},\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)p(c_i|\mathbf{c}_{-i})}{\sum_{c_i'}p\left(X_i^{(v)},\vec{F}_i^{(v)}|c_i',e_c,T^{(v)},\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)p(c_i'|\mathbf{c}_{-i})} \tag{A1}$$

The posterior probability of a particular value of c given the observed data, known transitivity categories, and other sentence category values is proportional to the likelihood, the probability of $X_i^{(v)}$ and $\vec{F}_i^{(v)}$ given that value of c, other observed data and category values, and the prior probability of c. We assume that c is independent of all other model parameters. The prior probability of c is a Dirichlet process (Ferguson 1973) with parameter α. In this process, each category value c_i has prior probability proportional to the number of sentence observations already assigned to that category, n_{c_i}. This process also reserves a small nonzero probability for new categories of c, determined by the parameter α, which we set equal to 1. The proportion of this extra probability that is reserved for new transitivity-violating categories is 0.19, the mean rate of transitivity violations inferred by the model in Perkins et al. 2022, and the proportion reserved for new categories without violations is set to 0.81. For n total observations of sentences across all categories, we define the prior on c.

$$p(c_i|\mathbf{c}_{-i})=\begin{cases}\dfrac{n_{c_i}}{n+\alpha} & \text{for previously seen values of } c\\ \dfrac{0.19\alpha}{n+\alpha} & \text{for new values where } e_c=1\\ \dfrac{0.81\alpha}{n+\alpha} & \text{for new values where } e_c=0\end{cases} \tag{A2}$$

Assuming independence between X and F, we calculate the likelihood as the product of the probabilities of observing $X_i^{(v)}$ and $\vec{F}_i^{(v)}$, given the other observations and model parameters.

$$p\left(X_i^{(v)},\vec{F}_i^{(v)}|c_i,e_c,T^{(v)},\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)=p\left(X_i^{(v)}|c_i,e_c,T^{(v)},\mathbf{X}_{-i},\mathbf{c}_{-i}\right)p\left(\vec{F}_i^{(v)}|c_i,e_c,\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right) \tag{A3}$$

The first term in this likelihood function is calculated differently depending on the value of e_c for the current category c_i. If c_i is a transitivity-violating category ($e_{c_i}=1$), then direct objects are generated by the grammatical property of that category $\delta_{c_i}^{(X)}$. We

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

calculate the probability of a direct object by integrating over all possible values of $\delta_{c_i}^{(X)}$, conditioning on other observations of sentences in this category.

$$p\left(X_i^{(v)}|c_i,e_i=1,T^{(v)},\mathbf{X}_{-i},\mathbf{c}_{-i}\right)=\int p\left(X_i^{(v)}|\delta_{c_i}^{(X)}\right)p\left(\delta_{c_i}^{(X)}|c_i,\mathbf{X}_{-i}\right)d\delta_{c_i}^{(X)} \tag{A4}$$

The first term inside the integral is equal to $\delta_{c_i}^{(X)}$ if $X_i^{(v)}=1$, or $1-\delta_{c_i}^{(X)}$ if $X_i^{(v)}=0$. We can use Bayes's rule to compute the second term inside the integral, the probability of $\delta_{c_i}^{(X)}$ given all other observations within the category.

$$p\left(\delta_{c_i}^{(X)}|c_i,\mathbf{X}_{-i}\right)=\frac{p\left(\mathbf{X}_{-i}|\delta_{c_i}^{(X)},c_i\right)p\left(\delta_{c_i}^{(X)}|c_i\right)}{\int p\left(\mathbf{X}_{-i}|\delta_{c_i}^{(X)},c_i\right)p\left(\delta_{c_i}^{(X)}|c_i\right)d\delta_{c_i}^{(X)}} \tag{A5}$$

The prior probability $p\left(\delta_{c_i}^{(X)}|c_i\right)$ is assumed to follow a uniform *Beta*(1,1) distribution. Let n_{c_i} be the total observations in category c_i and k_{c_i} be the total direct object observations in this category. The likelihood term, $p\left(\mathbf{X}_{-i}|\delta_{c_i}^{(X)},c_i\right)$, is the probability of observing k_{c_i} direct objects in n_{c_i} total observations. This follows a binomial distribution with parameter $\delta_{c_i}^{(X)}$.

$$p\left(\mathbf{X}_{-i}|\delta_{c_i}^{(X)},c_i\right)=\binom{n_{c_i}}{k_{c_i}}\left(\delta_{c_i}^{(X)}\right)^{k_{c_i}}\left(1-\delta_{c_i}^{(X)}\right)^{n_{c_i}-k_{c_i}} \tag{A6}$$

Solving the integral in equation A4, we calculate that $X_i^{(v)}$ takes a value of 1 with probability $\frac{k_{c_i}+1}{n_{c_i}+2}$, and 0 with probability $\frac{n_{c_i}-k_{c_i}+1}{n_{c_i}+2}$.

If c_i is not a transitivity-violating category ($e_{c_i}=0$), then direct objects in this category are generated by the transitivity properties of each verb. The first term in the likelihood function in equation A3 thus depends on the known transitivity category $T^{(v)}$ and $\theta^{(v)}$, the rate of direct objects under that transitivity category. If verb v is transitive or intransitive, then θ is known, and $X_i^{(v)}$ takes a value of 1 with probability θ, and 0 with probability $1-\theta$. If verb v is alternating, we again integrate over all possible values of $\theta^{(v)}$, conditioning on observations of this verb in other categories without argument gaps. This integral is analogous to the integral in equation A4. Here, let $n_1^{(v)}$ be the total observations for verb v in categories where $e_c=0$, and $k_1^{(v)}$ be the total direct object observations for verb v in these categories. Following equations analogous to A4–A6, we calculate that $X_i^{(v)}$ takes a value of 1 with probability $\frac{k_1^{(v)}+1}{n_1^{(v)}+2}$, and 0 with probability $\frac{n_1^{(v)}-k_1^{(v)}+1}{n_1^{(v)}+2}$.

The second term in equation A3 is the probability of the other observed features occurring in the given category. Assuming independence among features, this is equivalent to the product over the probabilities of observing each feature in this category.

$$p\left(\vec{F}_i^{(v)}|c_i,e_c,\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)=\prod_{F_i'^{(v)}}p\left(F_i'^{(v)}|c_i,e_c,\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right) \tag{A7}$$

The probability of observing a particular feature F in a category c_i is given by $\delta_{c_i}^{(F)}$ for that feature and that category. We integrate over all possible values of $\delta_{c_i}^{(F)}$, conditioning on other observations of feature F. Let n_{c_i} be the total observations in category c_i and $k_{c_i}^F$ be the total observations of feature F in this category. Following equations analogous to A4–A6, we calculate that $F_i^{(v)}$ takes a value of 1 with probability $\frac{k_{c_i}^F+1}{n_{c_i}+2}$, and 0 with probability $\frac{n_{c_i}-k_{c_i}^F+1}{n_{c_i}+2}$.

A1.2. Sampling e

After sampling values for c for each sentence in the data set, we then sample new values of e for each category. We calculate the posterior probability of each value of e_c for a category c given all of the direct object observations in that category $\mathbf{X}_c$ and known verb transitivity properties T.

$$p(e_c|c,\mathbf{X}_c,T)=\frac{p(\mathbf{X}_c|e_c,c,T)p(e_c)}{\sum_{e_c'}p\left(\mathbf{X}_c|e_c',c,T\right)p\left(e_c'\right)} \tag{A8}$$

We assume that e_c is independent of T and c and that the prior probability $p(e_c)=1$ is again set to 0.19, the mean rate of transitivity violations inferred by the model in Perkins et al. 2022. In other words, the learner assumes that the prior probability of a transitivity-violating category is equivalent to the probability that any single sentence contains a transitivity violation, as inferred by the previous learner. This will be the case only if sentences are equally distributed among categories, a simplifying assumption of the learner's prior that may be overridden if not supported by the data.

The likelihood term, $p(\mathbf{X}_c|e_c, c, T)$, is the probability of seeing particular observations of direct objects for verbs in this category. If $e_{c_i}=1$ and c_i is a transitivity-violating category, this probability is determined by $\delta_{c_i}^{(X)}$. We calculate the joint probability of the direct object observations for each verb in that category given $\delta_{c_i}^{(X)}$, integrating across all possible values of $\delta_{c_i}^{(X)}$.

$$p(\mathbf{X}_c|e_c=1,c,T)=\int\prod_{v'}\left(p\left(\mathbf{X}_c^{(v')}|\delta_{c_i}^{(X)}\right)\right)p\left(\delta_{c_i}^{(X)}|c_i\right)d\delta_{c_i}^{(X)} \tag{A9}$$

The first term inside the integral is the product across all verbs of probability of the direct observations for that verb $\mathbf{X}_c^{(v)}$ in the category, given $\delta_{c_i}^{(X)}$. This probability is given in equation A6. We again assume that the prior probability $p\left(\delta_{c_i}^{(X)}|c_i\right)$ follows a

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

uniform *Beta*(1,1) distribution. Let n_c be the total observations in a particular category and k_c be the total direct object observations in that category. Solving the integral in equation A9, we find the following.

$$p(\mathbf{X}_c|e_c=1,c,T)=\frac{\Gamma(k_c+1)\Gamma(n_c-k_c+1)}{\Gamma(n_c+2)}\left(\prod_{v'}\frac{\Gamma\left(n_c^{(v')}+1\right)}{\Gamma\left(k_c^{(v')}+1\right)\Gamma\left(n_c^{(v')}-k_c^{(v')}+1\right)}\right) \tag{A10}$$

If $e_{c_i}=0$ and c_i is not a transitivity-violating category, the likelihood term in equation A8 is determined by the known transitivity $T^{(v)}$ of each verb in the category. The probability of the particular direct object observations $\mathbf{X}_c$ in the category is the joint probability of seeing those direct object observations for each verb, given the transitivity of that verb.

$$p(\mathbf{X}_c|e_c=0,c,T)=\prod_{v'}\left(p\left(\mathbf{X}_c^{(v')}|T^{(v')}\right)\right) \tag{A11}$$

We can again rewrite $\mathbf{X}_c^{(v)}$ as k_c^v direct object observations out of n_c^v total observations for a given verb in a given category. The probability of observing k_c^v direct objects out of n_c^v total observations of a verb follows a binomial distribution with parameter $\theta^{(v)}$. Recall that $\theta^{(v)}=1$ for transitive verbs and $\theta^{(v)}=0$ for intransitive verbs. For alternating verbs, we must integrate across all possible values of $\theta^{(v)}$.

$$p(\mathbf{X}_c|e_c=0,c,T)=p\left(k_c^{(v)}|n_c^{(v)},T^{(v)}\right)=\int p\left(k_c^{(v)}|n_c^{(v)},\theta^{(v)}\right)p\left(\theta^{(v)}|T^{(v)}\right)d\theta^{(v)} \tag{A12}$$

We assume that $p(\theta^{(v)} \mid T^{(v)})$ follows a *Beta*(α,β) distribution, where the parameters α and β are counts of direct object observations and no direct object observations for verb v in other categories without argument gaps. Solving the integral in equation A12, we find as follows.

$$p\left(k_c^{(v)}|n_c^{(v)},T^{(v)}\right)=\left(\frac{\Gamma\left(n_c^{(v)}+1\right)}{\Gamma\left(k_c^{(v)}+1\right)\Gamma\left(n_c^{(v)}-k_c^{(v)}+1\right)}\right)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\left(\frac{\Gamma\left(k_c^{(v)}+\alpha\right)\Gamma\left(n_c^{(v)}-k_c^{(v)}+\beta\right)}{\Gamma\left(n_c^{(v)}+\alpha+\beta\right)}\right) \tag{A13}$$

A1.3. Sampling with annealing

The reported simulations used 5,000 total iterations of Gibbs sampling. This number was chosen to be the largest that could run within a feasible amount of time, and the model was run multiple times to assess convergence, with no substantive differences found across runs. To aid in the model's search process, simulated annealing was used during the first 1,000 iterations. In this process, we raise the posterior probabilities of c and e to the power of an annealing constant defined as $1/t$, where t is the current temperature. Then, we slowly lower the temperature (reduce t) until the annealing constant reaches 1. While the temperature is warm, the posterior probability distributions are flattened so the learner is able to explore more of its hypothesis space. After 1,000 iterations of Gibbs sampling with annealing, another 4,000 iterations were run without annealing. The final iteration was taken as a sample from the posterior distribution over c and e.

A2. Baseline models

A2.1. No-category baseline

Transitivity violations under the no-category baseline were sampled from the posterior probability distribution over the variable e in Perkins et al. 2022, given observed direct objects X and the values of T and $\delta^{(X)}$ inferred by that model. Here, e is a random variable encoding whether an individual sentence contains a transitivity violation and $\delta^{(X)}$ is the probability that a transitivity violation will produce a direct object in an individual sentence. Via Bayes's rule, the posterior predictive probability for the value e_i of a particular sentence i, given the direct object observation $X_i^{(v)}$ for the verb v in that sentence, all other error values $\mathbf{e}_{-i}$, other direct object observations $\mathbf{X}_{-i}$, and other model parameters, is as follows.

$$p\left(e_i|X_i^{(v)},T^{(v)},\delta^{(X)},\mathbf{e}_{-i},\mathbf{X}_{-i}\right)=\frac{p\left(X_i^{(v)}|e_i,T^{(v)},\delta^{(X)},\mathbf{e}_{-i},\mathbf{X}_{-i}\right)p(e_i)}{\sum_{e_i'}p\left(X_i^{(v)}|e_i',T^{(v)},\delta^{(X)},\mathbf{e}_{-i},\mathbf{X}_{-i}\right)p(e_i')} \tag{A14}$$

For the prior probability that a sentence contains a transitivity violation $p(e_i)$, we again use 0.19, the mean rate of transitivity violations (the parameter ε) inferred by the learner in Perkins et al. 2022. If the sentence contains a transitivity violation ($e_i=1$), the likelihood $p\left(X_i^{(v)}|e_i=1,T^{(v)},\varepsilon,\delta^{(X)},\mathbf{e}_{-i},\mathbf{X}_{-i}\right)$ depends only on the value for $\delta^{(X)}$, the probability that a transitivity violation produces a direct object: $X_i^{(v)}$ takes a value of 1 with probability $\delta^{(X)}$, and 0 with probability $1-\delta^{(X)}$. If the sentence does not contain a transitivity violation ($e_i=0$), the likelihood depends on the probability that verb v occurs with a direct object, given by $\theta^{(v)}$ for the verb's transitivity category $T^{(v)}$. If the verb is transitive or intransitive, $\theta^{(v)}$ is known; $X_i^{(v)}$ takes a value of 1 with probability θ, and 0 with probability $1-\theta$. If the verb is alternating, we must again integrate over all possible values of $\theta^{(v)}$, conditioning on other observations of this verb without transitivity violations. Let $n_1^{(v)}$ be the total observations for verb v in sentences where $e=0$, and $k_1^{(v)}$ be the total direct object observations in those sentences. Again following equations analogous to A4–A6, we find that $X_i^{(v)}$ takes a value of 1 with probability $\frac{k_1^{(v)}+1}{n_1^{(v)}+2}$, and 0 with probability $\frac{n_1^{(v)}-k_1^{(v)}+1}{n_1^{(v)}+2}$.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

For the no-category baseline simulation, values of e were randomly initialized for each sentence and were then resampled sequentially from the posterior distribution over e_i above, using the values for T and the mean value of $\delta^{(X)} = 0.25$ inferred by the model in Perkins et al. 2022. This process was repeated over 5,000 iterations of Gibbs sampling, and the final sample was used as a sample from model's posterior distribution over e.

A2.2. No-transitivity baseline

Sentence categories for the no-transitivity baseline were sampled according to a subset of the equations in Section A1 for our joint inference learner, without conditioning on individual verbs, transitivity values, or transitivity violations. Given observations of direct objects X_i and other features $\vec{F}_i$ in sentence i, together with other direct object and feature observations $\mathbf{X}_{-i}$ and $\mathbf{F}_{-i}$ as well as other sentence category assignments $\mathbf{c}_{-i}$, the posterior probability of c_i is as follows.

$$p\left(c_i|X_i,\vec{F}_i,\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)=\frac{p\left(X_i,\vec{F}_i|c_i,\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)p(c_i|\mathbf{c}_{-i})}{\sum_{c_i'}p\left(X_i,\vec{F}_i|c_i',\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)p(c_i'|\mathbf{c}_{-i})} \tag{A15}$$

The prior probability of c is calculated just as in our joint inference learner, in equation A2. Again assuming independence between X and F, we calculate the likelihood $p\left(X_i,\vec{F}_i|c_i,\mathbf{X}_{-i},\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)$ as the product of the likelihoods of observing X_i and $\vec{F}_i$ given all other observations and sentence categories. Because the no-transitivity baseline assumes that a direct object observation X_i is generated directly by the sentence category c_i and not by the transitivity properties of the verb in that sentence, the likelihood of a direct object $p(X_i|c_i,\mathbf{X}_{-i},\mathbf{c}_{-i})$ is calculated according to the equations for transitivity-violating categories under the joint inference learner (equations A4–A6). The likelihood of the features in a sentence $p\left(\vec{F}_i|c_i,\vec{\mathbf{F}}_{-i},\mathbf{c}_{-i}\right)$ is calculated in the same way as for our joint inference learner, in equation A7.

The simulation for the no-transitivity baseline was conducted similarly to the simulation for our joint inference learner. Each value of c was first randomly initialized to one of three categories and then resampled over 5,000 iterations of Gibbs sampling, with simulated annealing used during the first 1,000 iterations. The final iteration was taken as a sample from the posterior distribution over c.

Appendix B: Accuracy of automated data set coding

Table A1 reports percentage agreement and Cohen's kappa between two researchers' hand-coding and the automated annotations, for each sentence feature and clause type in a random sample of 500 sentences from our data set. We also report the interrater

	Rater 1 vs. Automated		Rater 2 vs. Automated		Rater 1 vs. Rater 2	
Feature	Agreement	Kappa	Agreement	Kappa	Agreement	Kappa
Subject overt	0.93	0.86	0.94	0.88	0.97	0.93
Subject sentence-initial	0.96	0.88	0.97	0.93	0.96	0.88
Subject follows aux	0.98	0.94	0.87	0.64	0.87	0.64
Subject follows NP	0.93	0.55	0.98	0.90	0.94	0.61
V is first in sentence	0.92	0.81	0.93	0.83	0.95	0.89
V before prep or particle	0.89	0.77	0.91	0.80	0.91	0.81
V has *-ed*	0.97	1.00	1.00	1.00	1.00	0.97
V has *-en*	1.00	1.00	1.00	1.00	1.00	1.00
V has *-ing*	1.00	1.00	1.00	1.00	1.00	1.00
V has *-s*	1.00	1.00	1.00	1.00	1.00	1.00
V irregular	1.00	0.96	1.00	1.00	1.00	0.96
V follows *to*	1.00	0.99	0.99	0.97	0.99	0.97
V follows *be*	0.98	0.91	0.92	0.71	0.94	0.80
V follows *have*	0.98	0.15	0.96	0.09	0.98	0.68
V follows *get*	1.00	0.66	0.99	0.44	0.99	0.72
V occurs with *do*	0.94	0.77	0.93	0.74	0.95	0.82
Sentence-initial functor	0.96	0.84	0.98	0.90	0.97	0.87
Sentence-medial functor pre-V	0.93	0.63	0.97	0.82	0.94	0.69
Sentence-medial functor post-V	0.93	0.34	0.95	0.41	0.95	0.61
Sentence-final functor	0.98	0.58	0.99	0.82	0.97	0.49
Question	1.00	0.99	1.00	1.00	0.99	0.99

Table A1. *Continued*

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

	Rater 1 vs. Automated		Rater 2 vs. Automated		Rater 1 vs. Rater 2	
Clause type						
Basic transitive	0.95	0.79	0.86	0.55	0.89	0.62
Basic intransitive	0.90	0.59	0.87	0.45	0.93	0.58
WH-question	0.94	0.72	0.94	0.66	0.97	0.87
Polar question	0.92	0.74	0.93	0.76	0.98	0.93
Other question	0.93	0.60	0.92	0.55	0.96	0.82
Passive	0.99	0.66	0.99	0.60	0.99	0.28
Relative clause	0.95	0.23	0.99	0.60	0.95	0.27
Other embedded clause	0.84	0.59	0.85	0.61	0.87	0.64
Imperative	0.95	0.80	0.95	0.75	0.96	0.82

Table A1. *Accuracy of sentence feature and clause-type coding.*

reliability of the two researchers for the purposes of comparison. We note that Cohen's kappa should be interpreted with caution, as each sentence feature and clause type was represented to different degrees within the 500-sentence sample.

Appendix C: Details of odds ratio comparisons

To assess the featural makeup of the categories inferred by our model, we calculated the odds ratio (OR) for each feature in each of these categories. The odds of observing a feature in a particular category were divided by the odds of observing the feature outside of that category. For any given category c and feature F, let $n_{F=1}^{(c)}$ be the number of times that feature was present within a category (had value 1) and $n_{F=1}^{(-c)}$ be the number of times that feature was present outside of that category. Similarly, let $n_{F=0}^{(c)}$ be the number of times that feature was absent within a category (had value 0) and $n_{F=0}^{(-c)}$ be the number of times that feature was absent outside of that category. Then, the odds ratio is calculated as in equation A16.

$$\mathrm{OR}_F^{(c)} = \frac{n_{F=1}^{(c)} \,/\, n_{F=1}^{(-c)}}{n_{F=0}^{(c)} \,/\, n_{F=0}^{(-c)}} \tag{A16}$$

An odds ratio greater than 1 indicates that a feature has higher-than-usual odds inside a category; an odds ratio less than 1 indicates that a feature has lower-than-usual odds inside a category. An odds ratio of infinity can occur if a feature is always present inside a category, and an odds ratio of 0 can occur if a feature is never present.

A Fisher's exact test was conducted to determine whether particular features had significantly higher or lower odds of occurring within a category. A Bonferroni correction was applied to correct for multiple comparisons: because twenty-two features were analyzed for each category (direct objects X plus all twenty-one features in $\vec{F}$), the critical value for each comparison was established by setting α equal to 0.05/22 = 0.002.

C1. Determining object-gap categories

We first determined our model's 'object gap' categories by calculating the odds of observing a direct object within each of the sixteen categories that the model inferred to have a transitivity violation. Table A2 reports the odds ratios (OR) for direct objects, along with their 95% confidence intervals (CIs) and p-values, in each of the model's transitivity-violating categories. A transitivity-violating category was classified as an object-gap category if the odds ratio was significantly lower than 1 (at $p <$ 0.002)—that is, if the category had significantly lower odds of producing direct objects. The threshold of significance was met in categories 1–15, which were therefore classified as object-gap categories; it was not met in category 16, which was therefore classified as 'other'.

We performed a similar calculation to determine the candidate object-gap categories for the no-transitivity baseline learner. Because this learner does not infer whether categories contain transitivity violations, we calculated the odds ratios for direct objects across all of the model's categories, reported in Table A3. A category was classified as a candidate object-gap category if it had significantly lower odds of producing direct objects (OR < 1, $p < 0.002$); this criterion was met in categories 1–22. All other categories were classified as 'no gap'.

C2. Analyzing features of object-gap categories

To assess which surface features F were distinctive of the joint inference model's fifteen object-gap categories, we calculated the odds ratios for each of the twenty-one features in each of these categories. Table A4 reports the odds ratio (OR), along with its 95%

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

Category	OR	CI	*p*-value
1	0.00	[0.00, 0.01]	4.97E-129
2	0.06	[0.03, 0.10]	8.70E-56
3	0.28	[0.22, 0.35]	1.71E-33
4	0.13	[0.06, 0.25]	2.19E-13
5	0.00	[0.00, 0.05]	5.47E-22
6	0.05	[0.00, 0.28]	3.35E-06
7	0.28	[0.14, 0.53]	1.44E-05
8	0.13	[0.06, 0.27]	2.11E-12
9	0.03	[0.00, 0.21]	3.64E-08
10	0.26	[0.13, 0.47]	5.36E-07
11	0.32	[0.26, 0.40]	4.76E-27
12	0.10	[0.06, 0.15]	8.04E-51
13	0.61	[0.48, 0.78]	4.01E-05
14	0.50	[0.37, 0.67]	2.01E-06
15	0.11	[0.06, 0.19]	6.91E-28
16	0.60	[0.39, 0.90]	0.01

Table A2. *Odds ratios for direct objects within transitivity-violating categories, joint inference model.*

Category	OR	CI	*p*-value	Category	OR	CI	*p*-value
1	0.10	[0.07, 0.14]	3.73E-85	23	1.61	[1.18, 2.20]	0.002
2	0.52	[0.43, 0.61]	9.17E-15	24	0.86	[0.58, 1.25]	0.41
3	0.27	[0.16, 0.44]	5.86E-09	25	0.64	[0.38, 1.04]	0.06
4	0.00	[0.00, 0.05]	2.79E-22	26	1.17	[0.79, 1.74]	0.44
5	0.00	[0.00, 0.19]	5.36E-07	27	57.05	[37.09, 92.92]	1.18E-292
6	0.13	[0.05, 0.27]	8.64E-12	28	1.12	[0.88, 1.43]	0.36
7	0.00	[0.00, 0.15]	8.98E-09	29	1.33	[1.00, 1.77]	0.05
8	0.14	[0.05, 0.32]	2.60E-08	30	5.16	[4.43, 6.03]	6.86E-130
9	0.12	[0.10, 0.14]	1.60E-172	31	0.41	[0.20, 0.81]	0.007
10	0.52	[0.45, 0.58]	8.39E-27	32	3.98	[2.98, 5.38]	5.41E-26
11	0.64	[0.55, 0.73]	2.95E-11	33	5.14	[3.92, 6.82]	2.33E-43
12	0.63	[0.51, 0.77]	8.01E-06	34	1.33	[1.22, 1.46]	6.45E-11
13	0.00	[0.00, 0.01]	8.90E-116	35	5.09	[4.46, 5.82]	2.75E-166
14	0.02	[0.00, 0.05]	6.66E-48	36	1.11	[0.92, 1.33]	0.27
15	0.04	[0.02, 0.08]	1.13E-67	37	1.09	[0.94, 1.25]	0.25
16	0.52	[0.44, 0.62]	3.78E-15	38	3.51	[2.69, 4.63]	2.10E-24
17	0.07	[0.04, 0.11]	4.77E-60	39	2.68	[2.29, 3.15]	1.61E-38
18	0.68	[0.55, 0.85]	0.0005				
19	0.41	[0.30, 0.56]	9.87E-10				
20	0.34	[0.25, 0.46]	2.28E-15				
21	0.25	[0.19, 0.32]	1.32E-32				
22	0.26	[0.13, 0.49]	2.29E-06				

Table A3. *Odds ratios for direct objects within sentence categories, no-transitivity baseline.*

confidence interval (CI) and *p*-value, for each feature in each of the model's object-gap categories. A feature was considered to be distinctive of a particular category if its odds ratio was significantly greater than 1 within that category ($p < 0.002$)—that is, if the category had significantly greater odds of producing that feature.

https://doi.org/10.1017/S0097850726000032 Published online by Cambridge University Press

Feature	OR	CI	*p*-value	OR	CI	*p*-value
		Category 1			Category 2	
Subject overt	∞	[79.90, ∞]	5.00E-97	0.06	[0.03, 0.09]	1.71E-74
Subject sent-init	0.00	[0, 0.02]	7.49E-56	0.00	[0.00, 0.04]	1.14E-34
Subject follows aux	1139.99	[325.67, 8192]	0	0.00	[0.00, 0.06]	2.27E-23
Subject follows NP	0.17	[0.08, 0.32]	2.91E-13	0.00	[0.00, 0.10]	3.22E-15
V is first in sentence	197.56	[35.30, 7456.10]	2.73E-68	0.05	[0.03, 0.07]	8.87E-96
V before prep or particle	0.45	[0.35, 0.56]	5.72E-14	0.64	[0.49, 0.84]	0.0008
V has *-ed*	0.00	[0.00, 0.26]	1.77E-06	0.00	[0.00, 0.43]	0.0003
V has *-en*	0.00	[0.00, 1.24]	0.13	0.00	[0.00, 2.02]	0.27
V has *-ing*	∞	[648.15, ∞]	0	0.00	[0.00, 0.06]	3.75E-23
V has *-s*	0.00	[0.00, 0.30]	1.39E-05	0.00	[0.00, 0.49]	0.001
V irregular	0.00	[0.00, 0.11]	2.38E-14	0.00	[0.00, 0.18]	4.62E-09
V follows *to*	0.00	[0.00, 0.03]	4.36E-43	383.75	[129.63, 1924.75]	1.16E-184
V follows *be*	∞	[907.56, ∞]	0	0.00	[0.00, 0.09]	1.14E-17
V follows *have*	0.00	[0.00, 0.84]	0.02	0.00	[0.00, 1.37]	0.12
V follows *get*	0.00	[0.00, 1.84]	0.27	0.00	[0.02, 3.00]	0.64
V occurs with *do*	0.00	[0.00, 0.05]	9.41E-28	0.00	[0.00, 0.09]	2.82E-17
Sent-init functor	∞	[828.6, ∞]	0	220.71	[109.42, 513.07]	1.96E-213
Sent-med functor pre-V	0.04	[0.01, 0.16]	1.85E-16	0.07	[0.01, 0.26]	1.58E-09
Sent-med functor post-V	0.00	[0.00, 0.31]	2.11E-05	0.00	[0.00, 0.51]	0.002
Sent-fin functor	0.30	[0.06, 0.89]	0.02	0.49	[0.10, 1.46]	0.29
Question	214.47	[72.76, 1044.77]	5.40E-162	389.07	[69.62, 13928.29]	3.27E-102
		Category 3			Category 4	
Subject overt	∞	[83.36, ∞]	3.95E-101	∞	[14.46, ∞]	1.22E-18
Subject sent-init	0.01	[0.00, 0.04]	4.52E-56	0.04	[0.00, 0.21]	1.95E-09
Subject follows aux	∞	674.71, ∞]	0	395.12	[68.96, 14149.22]	2.98E-61
Subject follows NP	0.47	[0.30, 0.69]	3.26E-05	1.97	[0.97, 3.14]	0.04
V is first in sentence	22.47	[11.72, 49.54]	4.87E-58	11.81	[3.90, 58.52]	9.51E-10
V before prep or particle	1.11	[0.91, 1.34]	0.29	0.45	[0.26, 0.75]	0.001
V has *-ed*	0.00	[0.00, 0.25]	1.26E-06	0.00	[0.00, 1.42]	0.19
V has *-en*	0.00	[0.00, 1.19]	0.08	0.00	[0.00, 6.75]	0.99
V has *-ing*	0.00	[0.00, 0.04]	9.44E-39	∞	[106.77, ∞]	4.30E-64
V has *-s*	0.00	[0.00, 0.29]	5.81E-06	0.00	[0.00, 1.63]	0.18
V irregular	0.00	[0.00, 0.11]	4.37E-15	0.00	[0.00, 0.60]	0.004
V follows *to*	0.02	[0.00, 0.06]	3.58E-41	0.00	[0.00, 0.18]	1.60E-08
V follows *be*	0.00	[0.00, 0.05]	1.01E-29	536.78	[93.64, 16384.00]	2.28E-70
V follows *have*	0.00	[0.00, 0.81]	0.03	0.00	[0.00, 4.57]	0.99

Table A4. *Continued*

Feature	OR	CI	*p*-value	OR	CI	*p*-value
V follows *get*	0.00	[0.00, 1.77]	0.28	0.00	[0.00, 10.01]	0.99
V occurs with *do*	40.09	[31.00, 52.51]	3.18E-264	0.00	[0.00, 0.30]	1.21E-05
Sent-init functor	54.78	[40.77, 74.90]	1.22E-286	0.23	[0.00, 0.68]	0.003
Sent-med functor pre-V	0.06	[0.01, 0.18]	4.50E-16	64.68	[34.02, 136.19]	4.63E-62
Sent-med functor post-V	0.92	[0.45, 1.68]	0.99	2.40	[0.75, 5.87]	0.07
Sent-fin functor	0.00	[0.00, 0.35]	6.37E-05	0.00	[0.00, 2.01]	0.27
Question	22.50	[15.32, 34.32]	2.04E-127	∞	[31.19, ∞]	1.85E-32
		Category 5			Category 6	
Subject overt	1.25	[0.76, 2.13]	0.41	0.00	[0.00, 0.11]	2.03E-10
Subject sent-init	1.75	[1.05, 2.89]	0.03	0.13	[0.00, 0.80]	0.02
Subject follows aux	1.32	[0.71, 2.34]	0.35	0.00	[0.00, 0.78]	0.01
Subject follows NP	0.67	[0.24, 1.54]	0.47	0.00	[0.00, 1.24]	0.10
V is first in sentence	0.00	[0.00, 0.02]	1.97E-38	0.31	[0.12, 0.76]	0.006
V before prep or particle	0.69	[0.40, 1.15]	0.15	0.58	[0.20, 1.47]	0.30
V has *-ed*	5.04	[2.29, 9.95]	8.06E-05	37.78	[15.54, 93.68]	3.85E-14
V has *-en*	43.00	[21.95, 79.61]	1.23E-18	0.00	[0.00, 24.72]	1.00
V has *-ing*	0.00	[0.00, 0.24]	1.31E-06	0.68	[0.13, 2.27]	0.79
V has *-s*	0.00	[0.00, 1.87]	0.27	0.00	[0.00, 5.99]	1.00
V irregular	19.54	[11.93, 32.36]	5.01E-31	5.46	[1.91, 13.89]	0.001
V follows *to*	0.36	[0.13, 0.82]	0.01	28.19	[8.41, 147.73]	2.11E-12
V follows *be*	0.00	[0.00, 0.33]	4.29E-05	147.16	[23.87, 5839.43]	2.33E-19
V follows *have*	0.00	[0.00, 5.22]	1.00	0.00	[0.00, 16.76]	1.00
V follows *get*	∞	[8402.05, ∞]	2.10E-192	0.00	[0.00, 36.70]	1.00
V occurs with *do*	0.00	[0.00, 0.34]	7.26E-05	0.00	[0.00, 1.09]	0.06
Sent-init functor	0.74	[0.31, 1.55]	0.50	0.26	[0.01, 1.62]	0.24
Sent-med functor pre-V	1.01	[0.39, 2.20]	1.00	0.86	[0.10, 3.52]	1.00
Sent-med functor post-V	0.51	[0.01, 2.98]	1.00	0.00	[0.00, 6.19]	1.00
Sent-fin functor	0.61	[0.02, 3.54]	1.00	1.92	[0.05, 11.88]	0.42
Question	1.23	[0.75, 1.99]	0.41	0.47	[0.15, 1.23]	0.15
		Category 7			Category 8	
Subject overt	2.20	[1.19, 4.32]	0.008	2.58	[1.47, 4.79]	0.0003
Subject sent-init	3.10	[1.81, 5.30]	1.86E-05	7.81	[4.69, 13.432]	1.97E-18
Subject follows aux	1.03	[0.49, 2.02]	0.87	0.25	[0.07, 0.67]	0.002
Subject follows NP	0.67	[0.21, 1.66]	0.55	0.00	[0.00, 0.36]	7.49E-05
V is first in sentence	13.05	[3.45, 110.38]	1.56E-07	5.38	[2.36, 15.16]	1.03E-06
V before prep or particle	0.80	[0.45, 1.38]	0.44	0.25	[0.12, 0.47]	8.10E-07
V has *-ed*	0.00	[0.00, 1.95]	0.27	0.00	[0.00, 1.49]	0.18

Table A4. *Continued*

Feature	OR	CI	*p*-value	OR	CI	*p*-value
V has *-en*	0.00	[0.00, 9.27]	1.00	9487.52	[2881.74, 4.50E+15]	1.44E-177
V has *-ing*	0.00	[0.00, 0.29]	1.19E-05	0.00	[0.00, 0.23]	5.91E-07
V has *-s*	0.00	[0.00, 2.25]	0.42	0.00	[0.00, 1.72]	0.28
V irregular	826.95	[142.55, 4.50E+15]	1.10E-68	0.00	[0.00, 0.63]	0.006
V follows *to*	0.00	[0.00, 0.25]	2.39E-06	0.00	[0.00, 0.19]	3.97E-08
V follows *be*	18.14	[10.04, 34.43]	1.66E-26	17.90	[10.68, 31.12]	1.32E-33
V follows *have*	177.85	[101.25, 317.74]	1.26E-57	660.85	[360.19, 1237.75]	9.35E-119
V follows *get*	0.00	[0.00, 13.74]	1.00	0.00	[0.00, 10.54]	1.00
V occurs with *do*	0.00	[0.00, 0.41]	0.0004	0.00	[0.00, 0.31]	3.12E-05
Sent-init functor	1.18	[0.54, 2.36]	0.58	0.87	[0.40, 1.71]	0.87
Sent-med functor pre-V	0.67	[0.18, 1.80]	0.66	0.64	[0.20, 1.57]	0.44
Sent-med functor post-V	0.62	[0.02, 3.60]	1.00	0.48	[0.01, 2.74]	0.73
Sent-fin functor	0.00	[0.00, 2.75]	0.65	0.00	[0.00, 2.11]	0.42
Question	0.50	[0.26, 0.90]	0.02	0.81	[0.50, 1.31]	0.42
		Category 9			Category 10	
Subject overt	∞	[5.18, ∞]	2.27E-07	1.93	[1.11, 3.50]	0.01
Subject sent-init	0.10	[0.00, 0.60]	0.003	4.90	[2.96, 8.23]	5.30E-11
Subject follows aux	1.93	[0.78, 4.36]	0.10	0.60	[0.25, 1.25]	0.21
Subject follows NP	70.82	[21.84, 362.20]	2.62E-23	0.82	[0.32, 1.79]	0.85
V is first in sentence	0.03	[0.00, 0.12]	2.63E-13	15.27	[4.07, 128.90]	5.52E-09
V before prep or particle	0.67	[0.28, 1.49]	0.36	1.83	[1.11, 3.02]	0.02
V has *-ed*	1.04	[0.03, 6.30]	0.62	587.73	[220.32, 2346.324]	1.38E-97
V has *-en*	136.01	[60.42, 303.36]	4.69E-23	0.00	[0.00, 7.92]	1.00
V has *-ing*	0.00	[0.00, 0.60]	0.004	0.00	[0.00, 0.25]	2.05E-06
V has *-s*	0.00	[0.00, 4.56]	1.00	0.00	[0.00, 1.92]	0.27
V irregular	5.43	[2.20, 12.31]	0.0002	0.00	[0.00, 0.70]	0.009
V follows *to*	0.96	[0.32, 2.40]	1.00	0.00	[0.00, 0.21]	2.45E-07
V follows *be*	8.84	[4.09, 19.64]	8.58E-09	3.26	[1.90, 5.45]	1.21E-05
V follows *have*	∞	[957.14, ∞]	4.92E-64	119.71	[70.98, 201.06]	3.62E-52
V follows *get*	0.00	[0.00, 27.90]	1.00	0.00	[0.00, 11.76]	1.00
V occurs with *do*	0.00	[0.00, 0.83]	0.03	0.00	[0.00, 0.35]	6.61E-05
Sent-init functor	0.00	[0.00, 0.76]	0.02	1.23	[0.60, 2.31]	0.50
Sent-med functor pre-V	2.78	[1.01, 6.65]	0.02	0.27	[0.03, 1.03]	0.06
Sent-med functor post-V	0.00	[0.00, 4.70]	1.00	0.00	[0.00, 1.98]	0.27
Sent-fin functor	1.47	[0.04, 8.90]	0.50	1.28	[0.15, 4.83]	0.67
Question	0.41	[0.15, 0.98]	0.04	0.97	[0.58, 1.60]	1.00

Table A4. *Continued*

Feature	OR	CI	*p*-value	OR	CI	*p*-value
		Category 11			Category 12	
Subject overt	0.09	[0.06, 0.11]	6.74E-101	98.97	[27.12, 810.97]	9.20E-62
Subject sent-init	0.02	[0.00, 0.06]	1.92E-50	0.06	[0.02, 0.13]	5.45E-29
Subject follows aux	0.00	[0.00, 0.04]	1.37E-37	0.08	[0.03, 0.19]	1.66E-18
Subject follows NP	0.00	[0.00, 0.06]	1.22E-24	112.56	[71.80, 186.11]	3.12E-233
V is first in sentence	0.12	[0.09, 0.15]	8.62E-94	0.28	[0.22, 0.35]	2.43E-27
V before prep or particle	0.66	[0.53, 0.813]	3.62E-05	1.13	[0.89, 1.43]	0.31
V has *-ed*	0.00	[0.00, 0.26]	1.81E-06	5.32	[3.69, 7.51]	8.41E-16
V has *-en*	0.67	[0.08, 2.48]	1.00	0.00	[0.00, 1.83]	0.27
V has *-ing*	0.00	[0.00, 0.04]	2.29E-37	0.00	[0.00, 0.06]	1.77E-254
V has *-s*	0.08	[0.00, 0.45]	0.0001	5.10	[3.43, 7.37]	2.46E-13
V irregular	0.15	[0.05, 0.35]	7.00E-09	4.59	[3.46, 6.03]	2.38E-22
V follows *to*	278.48	[133.83, 704.94]	3.82E-297	0.57	[0.39, 0.80]	0.0007
V follows *be*	0.01	[0.00, 0.08]	8.22E-27	0.09	[0.02, 0.22]	3.71E-14
V follows *have*	0.00	[0.00, 0.83]	0.02	0.00	[0.00, 1.24]	0.13
V follows *get*	3.16	[1.12, 7.25]	0.02	2.26	[0.45, 6.89]	0.16
V occurs with *do*	0.01	[0.00, 0.08]	5.63E-26	0.11	[0.04, 0.26]	1.22E-12
Sent-init functor	0.50	[0.34, 0.71]	3.96E-05	0.34	[0.19, 0.56]	1.41E-06
Sent-med functor pre-V	8.27	[6.77, 10.11]	1.54E-82	18.48	[14.43, 23.77]	8.62E-119
Sent-med functor post-V	1.81	[1.08, 2.87]	0.01	2.45	[1.42, 3.99]	0.001
Sent-fin functor	0.50	[0.16, 1.19]	0.14	0.00	[0.00, 0.54]	0.002
Question	0.65	[0.53, 0.80]	2.93E-05	0.66	[0.51, 0.84]	0.0007
		Category 13			Category 14	
Subject overt	0.03	[0.02, 0.05]	3.68E-96	67.12	[18.33, 557.28]	
Subject sent-init	0.04	[0.01, 0.10]	1.12E-31	36.47	[21.84, 65.11]	1.93E-41
Subject follows aux	0.02	[0.00, 0.09]	6.37E-24	0.07	[0.01, 0.21]	2.30E-93
Subject follows NP	0.00	[0.00, 0.09]	6.84E-17	1.24	[0.81, 1.86]	1.51E-13
V is first in sentence	2.80	[2.00, 4.00]	1.49E-11	5.62	[3.32, 10.24]	0.27
V before prep or particle	0.52	[0.40, 0.68]	2.88E-07	0.60	[0.44, 0.81]	5.40E-16
V has *-ed*	0.53	[0.17, 1.26]	0.18	2.71	[1.51, 4.56]	0.0007
V has *-en*	0.99	[0.12, 3.70]	1.00	2.22	[0.45, 6.75]	0.0006
V has *-ing*	0.59	[0.39, 0.85]	0.003	0.12	[0.04, 0.28]	0.16
V has *-s*	0.24	[0.03, 0.88]	0.03	3.58	[2.06, 5.87]	2.20E-11
V irregular	0.90	[0.53, 1.44]	0.73	2.04	[1.30, 3.08]	0.001
V follows *to*	0.01	[0.00, 0.07]	1.33E-27	0.08	[0.02, 0.21]	7.71E-15
V follows *be*	0.00	[0.00, 0.08]	1.46E-19	0.00	[0.00, 0.11]	1.28E-13
V follows *have*	0.00	[0.00, 1.24]	0.13	0.49	[0.01, 2.80]	0.73

Table A4. *Continued*

Feature	OR	CI	*p*-value	OR	CI	*p*-value
V follows *get*	0.00	[0.00, 2.71]	0.65	1.07	[0.03, 6.18]	0.61
V occurs with *do*	0.84	[0.57, 1.21]	0.39	0.55	[0.31, 0.93]	0.02
Sent-init functor	0.14	[0.06, 0.30]	3.60E-12	0.41	[0.21, 0.72]	0.0006
Sent-med functor pre-V	0.00	[0.00, 0.12]	1.58E-13	0.38	[0.16, 0.78]	0.004
Sent-med functor post-V	0.25	[0.03, 0.91]	0.03	3.24	[1.80, 5.45]	9.95E-05
Sent-fin functor	1.22	[0.52, 2.46]	0.55	4.14	[2.34, 6.90]	3.18E-06
Question	0.26	[0.18, 0.35]	6.98E-22	0.51	[0.37, 0.71]	1.60E-05
		Category 15				
Subject overt	113.66	[20.05, 4391.44]	2.82E-36			
Subject sent-init	0.76	[0.51, 1.11]	0.16			
Subject follows aux	0.76	[0.47, 1.18]	0.27			
Subject follows NP	13.25	[9.60, 18.41]	2.79E-57			
V is first in sentence	0.80	[0.58, 1.11]	0.18			
V before prep or particle	0.59	[0.42, 0.83]	0.001			
V has *-ed*	0.00	[0.00, 0.67]	0.007			
V has *-en*	0.00	[0.00, 3.20]	0.63			
V has *-ing*	∞	[232.04, ∞]	6.42E-133			
V has *-s*	0.00	[0.00, 0.78]	0.02			
V irregular	0.00	[0.00, 0.28]	6.65E-06			
V follows *to*	0.00	[0.00, 0.09]	3.49E-17			
V follows *be*	65.85	[39.18, 118.18]	1.86E-116			
V follows *have*	0.00	[0.00, 2.17]	0.42			
V follows *get*	0.00	[0.00, 4.76]	1.00			
V occurs with *do*	0.00	[0.00, 0.14]	3.57E-11			
Sent-init functor	1.33	[0.87, 1.98]	0.15			
Sent-med functor pre-V	9.69	[7.07, 13.28]	2.45E-41			
Sent-med functor post-V	1.58	[0.62, 3.36]	0.22			
Sent-fin functor	0.00	[0.00, 0.95]	0.04			
Question	1.38	[1.01, 1.88]	0.04			

Table A4. *Odds ratios for features* F *within object-gap categories, joint inference model.*

Laurel Perkins is an assistant professor in the Department of Linguistics at the University of California, Los Angeles. Email: perkinsl@ucla.edu.

Naomi H. Feldman is a professor in the Department of Linguistics and the Institute for Advanced Computer Studies at the University of Maryland. Email: nhf@umd.edu.

Jeffrey Lidz is a professor in the Department of Linguistics at the University of Maryland. Email: jlidz@umd.edu.

Cite this article: Perkins, Laurel; Naomi H. Feldman; and Jeffrey Lidz. 2026. Mind the gap: Learning the surface forms of movement dependencies. *Language* 1–42. https://doi.org/10.1017/S0097850726000032.