Journal of Memory and Language 145 (2025) 104672

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Reframing linguistic bootstrapping as joint inference using visually-grounded grammar induction models

Eva Portelance [a,b,*,1], Siva Reddy [c,d,b,2], Timothy J. O'Donnell [d,b,3]

[a] Department of Decision Sciences, HEC Montréal, Montreal, Canada
[b] Mila - Quebec AI Institute, Canada
[c] Department of Computer Science, McGill University, Montreal, Canada
[d] Department of Linguistics, McGill University, Montreal, Canada

ARTICLE INFO

Keywords:
Syntactic bootstrapping
Semantic bootstrapping
Multimodal models
Grammar induction
Word learning
Language models
Cognitive models

ABSTRACT

Semantic and syntactic bootstrapping posit that children use their prior knowledge of one linguistic domain, say syntactic relations, to help later acquire another, such as the meanings of new words. Empirical results supporting both theories may tempt us to believe that these are different independent learning strategies. Here, we argue for a unified approach, where instead they are both contingent on a more general learning strategy for language acquisition: joint learning. Using a series of neural visually-grounded grammar induction models, we demonstrate that both syntactic and semantic bootstrapping effects are strongest when syntax and semantics are learnt simultaneously via joint learning. This more general learning strategy results in better grammar induction, realistic lexical category learning, and better interpretations of novel sentence and verb meanings. Joint learning makes language acquisition *easier* for learners by mutually constraining the hypotheses spaces for both syntax and semantics. Studying the dynamics of joint inference over many input sources and modalities represents an important new direction for language modeling and learning research in both cognitive sciences and AI, as it may help us explain how language can be acquired in more constrained learning settings.

Introduction

With the advent of large language models (LLMs) it has become hard to deny that, if you have little to no limitations on the quantity of input data, amount of computation, or memory size, much linguistic structure can be learnt (Mahowald et al., 2024; Piantadosi, 2023). Clearly human learners are not as unconstrained and therefore LLMs cannot be said to mimic human language learning. However, in light of the fact that these models can learn what they do without the need for language-specific learning mechanisms or innate linguistic knowledge, it is important to reassess existing debates on language acquisition in children.

Language models have been proposed as useful tools for building proof-of-concept arguments for what is learnable from linguistic input (Lappin, 2021, ch. 1.2, Pearl, 2023; Portelance & Jasbi, 2024; Tsuji, Cristia, & Dupoux, 2021; Warstadt & Bowman, 2023). As such, they can inform innateness debates (Clark & Lappin, 2011; Crain & Thornton, 2012), which try to establish how much innate knowledge is necessary to learn language and how specific to language learning strategies must be for its acquisition. Proofs of concept are however limited to informing us of what can possibly be learnt under the learning conditions of the target model. Portelance and Jasbi (2024) suggest that a more interesting way to use these language models is for hypothesis generation, whereby models are used to study the dynamics at play during language learning to propose novel theories about the strategies which drive language acquisition in people. This paper adopts this approach, proposing a unified explanation for linguistic bootstrapping phenomena.

Linguistic bootstrapping theories posit that children use their prior knowledge in one linguistic domain, for example syntactic relations, to help with the acquisition of another, such as the meanings of new words. Here, we will address two theories, semantic bootstrapping and syntactic bootstrapping. These proposals come from a theoretical landscape which assumed that learning is algorithmic or stage-based in nature, requiring some prior innate linguistic knowledge as a starting point. The bootstrapping debates have centered on the questions *what linguistic knowledge do we start from?* and *how do children use*

* Corresponding author at: Department of Decision Sciences, HEC Montréal, Montreal, Canada.
E-mail address: eva.portelance@hec.ca (E. Portelance).
[1] IVADO Professor.
[2] Facebook CIFAR AI Chair.
[3] Canada CIFAR AI Chair.

https://doi.org/10.1016/j.jml.2025.104672
Received 1 November 2023; Received in revised form 23 June 2025; Accepted 25 June 2025
Available online 24 July 2025

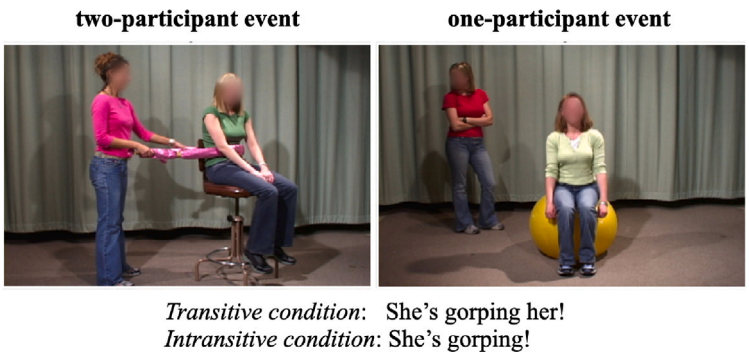


Fig. 1. Example of nonce verb learning experimental paradigm from (Yuan, Fisher, & Snedeker, 2012), demonstrating indirect evidence for the syntactic bootstrapping hypothesis.

this knowledge to bootstrap new knowledge and eventually acquire language? (Gleitman, 1990; Grimshaw, 1981; Landau & Gleitman, 1985; Pinker, 1984).

Linguistic bootstrapping debates

Though linguistic bootstrapping can extend to all levels of linguistic representation, from phonology to pragmatics, we will concentrate on the two types which have been most discussed, semantic bootstrapping and syntactic bootstrapping.

Broadly, in semantic bootstrapping proposals, children are said to use their knowledge of semantics and meaning to *bootstrap* syntactic knowledge. The most famous formulation comes from Pinker (1984, 2009). In this version, semantic bootstrapping is envisioned as an early language learning strategy which allows children to learn syntactic primitives such as syntactic categories (noun, verb, adjective...), by mapping them to early perceptual or cognitive categories[4] (individual, action, state...). For example, children may induce a syntactic category like noun by noticing that there are words which name the semantic perceptual categories of persons or things. Once categories are learnt, a more symbiotic exchange would then emerge between syntactic and semantic knowledge to acquire new structures and word meanings. There is indirect empirical evidence which supports this hypothesis, such as the observation that children's first nouns correspond to physical objects, first verbs to actions, and first adjectives to perceptually salient attributes. Pinker describes semantic bootstrapping as a form of distributional learning, stating that "children always give priority to distributionally based analyses, and [semantic bootstrapping] is intended to explain how the child knows which distributional contexts are the relevant ones to examine" (Pinker, 2009, p. 42–43). In this view, language learning can be thought of as a probabilistic mapping problem. Semantic bootstrapping is then a theory that tries to explain how children use aspects of meaning to acquire core pieces of syntactic structure.

Syntactic bootstrapping proposals instead describe processes which involve the use of syntactic knowledge to *bootstrap* new meanings and semantic knowledge. The proposal and first full description of this theory is associated with Gleitman (1990). Gleitman makes the case that the same problems which plague syntax acquisition and have motivated many syntactic theories, also exist for vocabulary acquisition: the hypothesis space over syntactic structure and word meanings is simply too vast and must be limited in some way by prior knowledge or learning strategies to account for children's language learning abilities. As a way to limit this hypothesis space for vocabulary acquisition, she initially posits that learners must start with "sophisticated presuppositions about the structure of language". In other words, children start with some syntactic knowledge to bias their acquisition of new word meanings. Discussion of syntactic bootstrapping have for the most part been limited to the acquisition of verb meanings, though in principle it may extend to other types of 'hard' words (Fisher, Gertner, Scott, & Yuan, 2010; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). There is once again indirect empirical evidence which supports this hypothesis, such as children's ability to infer a novel verb's meaning based on its argument structure (Naigles, 1990). For example, if a child is exposed to the transitive condition in Fig. 1 and they have learnt to identify such argument structures, they can infer the presence of an agent (someone performing an action) and a patient (someone upon whom an action is performed) are likely necessary in the correct visual interpretation, picking the two-participant event—and more generally identifying and learning the meaning of 'gorping'. In its original strong form, syntactic bootstrapping requires innate language-specific abstract knowledge of syntactic structures, which is independent of word meaning or grounding. More recent descriptions however present early abstract structural knowledge as possibly probabilistic and learnt from more primitive concepts (Fisher, Jin, & Scott, 2020; Gleitman et al., 2005).

Though sometimes erroneously characterized as opposing theories in the mythology of the field, semantic and syntactic bootstrapping are complementary theories.[5] Their characterization as conflicting may stem from the conjecture that semantic bootstrapping leads to a lexically-driven view of language acquisition where language is learnt by mapping words to meanings and syntactic primitives (bottom-up perspective), while syntactic bootstrapping leads to a more abstraction-based view where language is learnt by mapping sentential structures to word meanings (top-down perspective). However, these theories can very much coexist. In Pinker (2009)'s semantic bootstrapping proposal, the second stage of learning after primitive syntactic categories are learnt is called "structure-dependent distributional learning" and resembles syntactic bootstrapping. In Gleitman et al. (2005), syntactic bootstrapping is also characterized as a multi-stage process where, initially children learn words – mostly concrete nouns – by mappings them to salient referents in their extra-linguistic context. Once enough of these words are known they can infer syntactic knowledge such as clausal structure to bootstrap the acquisition of harder words — such as verbs or adjectives. In either case, one theory does not exclude the other; semantic and syntactic bootstrapping coexistence is possible, they are simply studied as independent learning strategies, where in principle the other strategy could also be at work. In their review article on syntactic bootstrapping, Fisher et al. (2020) articulate the current leading view: that these are independent learning strategies which both rely on the same tight bond between syntactic and semantic representation.

[4] Somewhat confusingly, it is called *semantic* bootstrapping, even though in its original formulation, it never refers to any knowledge of formal semantics or linguistic meaning, but instead to perceptual concepts independent of language.

[5] See footnote 4 of Gleitman et al. (2005) for discussion of possible origins of this misconception.

Grammar induction models and linguistic bootstrapping

Grammar induction is the task of learning a grammar – or set of rules and structures – given a corpus of sentences. In the past, it had proven quite difficult to do using purely statistical learning or distributed models, especially for natural language which requires more expressive grammars, such as probabilistic context-free grammars (PCFG; Cohn, Blunsom, & Goldwater, 2010; Klein & Manning, 2004, 2005; Perfors, Tenenbaum, & Regier, 2011). At the time, the solution had been to use probabilistic models with more informed prior knowledge of language, such as access to categories, head-branching bias or semantic mappings (Chater & Manning, 2006; Muralidaran, Spasić, & Knight, 2020). One such model of particular interest to this paper was the semantic bootstrapping model from Abend, Kwiatkowski, Smith, Goldwater, and Steedman (2017), which was an instantiation of the theory by the same name. The authors proposed a Bayesian grammar induction model which learnt both syntactic derivations for sentences and word meanings conditioned on knowing sentential meanings (given in the form of compositional semantic derivations). Using their model, the authors argued that syntactic bootstrapping is an emergent effect which follows from this tight bond imposed on syntax and semantics during semantic bootstrapping. In such a proposal, syntactic bootstrapping is not a learning strategy in its own right, but an effect which follows from semantic bootstrapping strategies.

More recently, access to better computational resources and neural network architectures have lead to significant advances for grammar induction models – as with language models. Researchers have managed to design successful distributional approaches to natural language grammar induction without the need for overly informative priors (Drozdov, Verga, Yadav, Iyyer, & McCallum, 2019; Kim, Dyer, & Rush, 2019). These models have since been augmented by introducing visual-grounding, finding that access to visual information helps models induce grammars producing more accurate constituency trees (Jin & Schuler, 2020; Shi, Mao, Gimpel, & Livescu, 2019; Wan, Han, Zheng, & Tuytelaars, 2022; Zhao & Titov, 2020) — a result reminiscent of semantic bootstrapping.

Neural grammar induction models are different from neural language models in that they explicitly enforce the learning of a grammar that follows a set of formal constraints, for example that the grammar be a PCFG. This approach to grammar and language learning is different from neural language models such as LLMs or even smaller language models trained on child directed utterances (Huebner, Sulem, Cynthia, & Roth, 2021) which generally make fewer assumptions about the structure of language and in turn do not explicitly learn a grammar, but may implicitly do so. Given their implicit nature, the most we can do with LLMs is to study their external linguistic productions with carefully designed probes to see what grammatical knowledge they may have acquired (Hu, Gauthier, Qian, Wilcox, & Levy, 2020; Linzen, Dupoux, & Goldberg, 2016; Warstadt et al., 2020) – in other words they are observationally adequate models of language learning. Neural grammar induction models on the other hand have explicit representations of grammar which can be directly studied, affording them a greater degree of descriptive adequacy (Portelance & Jasbi, 2025). Bootstrapping theories are about the internal learning procedures that lead to specific types of grammatical and semantic knowledge, for example syntactic categories or semantic roles. With this paper, we want to go beyond extrinsic evidence and seek intrinsic evidence.

Taking inspiration from recent advances in neural grammar induction, we present our own model, building on previous work, which additionally learns to interpret visual information via access to induced trees. Previous work on visually-grounded grammar induction used fixed pretrained image embeddings as their representations for visual semantic context, while we train our visual embeddings from scratch such that induced grammars impact updates to these representations throughout learning. Thus, in addition to semantic bootstrapping, our model has the possibility of performing syntactic bootstrapping by allowing induced syntactic knowledge to affect the acquisition of semantic representations. Like Abend et al. (2017), our model addresses both semantic and syntactic acquisition; however being a neural network, we are able to introduce the additional complication of learning not just grammar, but semantic representations for words and sentences from scratch as well. It thus more closely models young children's reality. We will use this model to make a different proposal about the relation between syntactic and semantic bootstrapping from Abend et al. (2017), who argued that syntactic bootstrapping is a downstream effect of semantic bootstrapping as the main learning strategy. We propose that syntactic and semantic bootstrapping are both downstream effects of a more general joint learning strategy. While still honoring the leading view that syntactic and semantic bootstrapping are complementary acquisition strategies (Fisher et al., 2020), we differ from these proposals in that we do not study them as independent learning processes, but instead offer a unified account relying on their interdependence.

Our proposal

In this paper, we make the following theoretical proposal: linguistic bootstrapping follows from joint learning over multiple levels of linguistic representation, via simultaneous access to multiple input modalities. We argue that neither syntactic nor semantic bootstrapping are independent learning strategies, as they have previously been presented, but both learning effects which arise as a consequence of a probabilistic joint learning strategy over both syntactic and semantic levels of representation for language, highlighting instead their interdependence. In other words, these bootstrapping effects simply arise from wanting to learn word meanings and language structure at the same time. Furthermore, we propose that no prior linguistic knowledge is necessary beyond a bias towards learning abstract categories (syntactic or conceptual), to acquire both syntactic and semantic representations.

Learning syntactic structures can facilitate learning word and sentence meanings, and conversely, learning meanings can facilitate learning syntactic structures. In a joint inference process over both syntactic and semantic representations, each hypothesis space can constrain the other and help learners to simultaneously acquire syntax and semantics.

In the sections which follow, we will use a neural visual grammar induction model to show how constraints during joint inference on one linguistic domain can affect another and lead to better generalization in completely novel contexts.[6] Our model learns grounded representations of both syntactic structure and semantic meanings from sentence-image pairs using a statistical learning algorithm. We ask the following questions: **1.** (analogous to semantic bootstrapping) Can access to visual-grounding and the ability to learn semantic representations in a joint learning setting facilitate learning grammars that generalize better to unseen contexts? **2.** (analogous to syntactic bootstrapping) Can access to linguistic structure and the ability to learn grammar in a joint learning setting facilitate learning and interpreting novel words and contexts?

The dataset

For all of the demonstrations and experiments reported in this paper, we use the Abstract Scenes dataset (Zitnick & Parikh, 2013; Zitnick, Parikh, & Vanderwende, 2013), a dataset composed of clip-art scenes meant to resemble children's book illustrations paired with simple sentences narrating the contents of the images, Fig. 2 contains examples of these image-sentence pairs.

Previous work on visually-grounded grammar induction has used image-caption pairs from datasets like MS-COCO (Chen et al., 2015;

[6] All data, code, analyses and experimental results are publicly available at [Anonymized].

Fig. 2. Examples image-sentence pairs from the Abstract Scenes dataset (Zitnick & Parikh, 2013; Zitnick et al., 2013).

Lin et al., 2014); the problem with image captions is that they are not always complete sentences, but often complex noun phrase descriptions which lack main verbs, for example "A man sitting on a park bench with an umbrella". For this work, it was important to use image-sentence pairs with complete sentences containing main verbs. Much of the existing literature on linguistic bootstrapping, especially syntactic bootstrapping, has centered on the acquisition of verb meaning using argument structure. Furthermore, not having access to main verbs enough may bias grammar induction models towards trees that favor different words as sentence heads.

The images and sentences from Abstract Scenes contain depictions and descriptions of actions with both transitive and intransitive main verbs that lend themselves well to syntactic and semantic bootstrapping experimentation. This dataset has also been used before in modeling experiments for language acquisition research (Nikolaus & Fourtassi, 2021a, 2021b). In total it contains 10,020 images each paired with 6 different sentences, for a total of 60,160 image-sentence pairs.

Test–train splits

The Abstract Scenes Dataset does not come with preexisting data splits, so we designed our own to evaluate syntactic and semantic bootstrapping on in-distribution syntactic structure learning and out-of-distribution verb learning.

For the out-of-distribution experiment presented in Section "Experiment 2: Syntactic bootstrapping and joint learning", we created a test split that contained image-sentence pairs with novel main verbs. In other words, if the model was trained on instances of *throw, kiss, cry* it was then evaluated on *toss, hug, smile*. The latter verbs can be considered nonce verbs from the perspective of the model. To create our test data, we first extracted all of the main verbs with more than 5 instances in the dataset (around 480 verbs), grouped them by verb stem (280 unique verb stems), annotated them for transitivity and object animacy. We then hand selected 10 intransitive verb stems and 10 transitive stems taking animate objects and 10 taking inanimate objects. We selected these verbs such that they appeared in varied sentential contexts (i.e. there were no synomyms or closely related verbs) and such that the total number of transitive and intransitive sentences were as close as possible. The training and test verb stem lists are available in Appendix "Verb stem lists for data split". In total, the test set contained 1708 sentences (718 transitive, 990 intransitive) with 30 different held out verb stems (10 transitive-animate, 10 transitive-inanimate, 10 intransitive).

For experiments in Section "Experiment 1: Semantic bootstrapping and joint learning" with in-distribution tests, we simply reintroduced half of the test set to training, so that all verbs were now seen at least once during training, resulting in 833 test sentences (357 transitive, 476 intransitive) with 30 different verb stems. Thus, test sentences in these in-distribution evaluation were novel sentences containing known verbs.

The joint-learning model

Our joint-learning model must learn both syntactic and semantic knowledge from grounded linguistic input. As mentioned in the introduction, a family of models which lend themselves well to learning explicit grammatical representations are neural grammar induction models. In particular, since we would like to eventually have syntactic and semantic representations of acquired knowledge that can effect one another through bootstrapping-like mechanisms during learning, we need our grammar induction model to be semantically grounded in some way. For this reason, we base our model on a visually-grounded grammar induction model called VC-PCFG (visually-grounded compound probabilistic context-free grammar) by Zhao and Titov (2020). Our joint-learning model combines two types of learning objectives together: (1) a syntactic learning objective and (2) a semantic learning objective, respectively trying to learn syntactic and semantic knowledge from visually grounded linguistic input. We describe each of these objective in the subsections which follow.

The syntactic objective

We will represent syntactic knowledge as a compound probabilistic context free grammar, or C-PCFG (Kim et al., 2019). C-PCFGs are extensions of probabilistic context free grammars (PCFGs). A PCFG, $\langle \mathcal{G}, \boldsymbol{\pi} \rangle$ is a grammar $\mathcal{G}$ coupled with $\boldsymbol{\pi}$, a probability function over $\mathcal{G}$. Here, the context free grammar can be formalized as a 5-tuple of finite sets $\mathcal{G} = (S, \mathcal{N}, \mathcal{P}, \Sigma, \mathcal{R})$, where S is the start symbol, $\mathcal{N}$ the nonterminal categories, $\mathcal{P}$ the preterminal categories, Σ the vocabulary or set of terminals, and $\mathcal{R}$ the production rules over words and categories. The rules in $\mathcal{R}$ are in Chomsky normal form:

$$\begin{aligned} S &\to A, & A &\in \mathcal{N}, \\ A &\to BC, & A &\in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P}, \\ T &\to w, & T &\in \mathcal{P}, w \in \Sigma. \end{aligned}$$

The probability function $\boldsymbol{\pi}$ assigns some non-negative value to every production rule $r \in \mathcal{R}$. In most PCFGs, it is defined as a set of categorical probability distributions where there is a separate categorical for every set of rules with the same left-hand side such that $\sum_{\alpha | A \to \alpha} \pi(A \to \alpha) = 1$. However, as shown in Kim et al. (2019), the strong context-free assumption instantiated by this type of probability function is not conducive to effectively inducing a grammar from scratch. Instead, we need a way to share information across rule applications in a tree. The proposed solution in Kim et al. (2019) is the C-PCFG, which assumes that rule probabilities $\boldsymbol{\pi}$ follow a compound distribution (Robbins, 1951). In such a distribution we additionally condition each rule probability on some contextual information, here a latent representation of the complete sentence's meaning that we will call $\mathbf{z}$. Intuitively, $\mathbf{z}$ represents shared information in each sentence that is accessible throughout its tree derivations at each rule application. C-PCFGs rule probabilities are thus not strictly independent of one another. C-PCFGs

are in fact a mixture of PCFGs, satisfying the context-free assumption when conditioned on some value for the random variable $\mathbf{z}$, here an abstract representation for the overall meaning of a sentence. Access to additional information in the form of this sentence-level representation turns out to be crucial for successful grammar induction,[7] making what used to be a very hard problem – inducing a grammar from scratch – solvable. In practice, we will sample a $\mathbf{z}$ from a spherical Gaussian distribution with prior parameters γ for each sentence in our input.

$$\mathbf{z} \sim \text{SphericalGaussian}(\gamma), \tag{1}$$

Additionally, in a neural C-PCFG we use embeddings rather than symbols such as "VP" or "N" to represent all our possible left and right-hand sides of rules. Thus, we must parameterize the probability of each individual rule using embeddings, where $\mathbf{u}, \mathbf{w}$ will represent embeddings for possible left and right-hand sides of rules respectively. Our C-PCFG, $\langle \mathcal{G}, \boldsymbol{\pi} \rangle$, defines a compound probability distribution over rules using the following process (note that all rule probabilities are conditioned on $\mathbf{z}$):

$$\pi_{\mathbf{z}}(r) = \begin{cases} \frac{\exp(\mathbf{u}_A^\top f_s([\mathbf{w}_S;\mathbf{z}]))}{\sum_{A' \in \mathcal{N}} \exp(\mathbf{u}_{A'}^\top f_s([\mathbf{w}_S;\mathbf{z}]))}, & \text{for } r \in S \to A \\ \frac{\exp(\mathbf{u}_{BC}^\top [\mathbf{w}_A;\mathbf{z}])}{\sum_{B',C' \in \mathcal{N} \cup P} \exp(\mathbf{u}_{B'C'}^\top [\mathbf{w}_A;\mathbf{z}])}, & \text{for } r \in A \to BC \\ \frac{\exp(\mathbf{u}_w^\top f_t([\mathbf{w}_T;\mathbf{z}]))}{\sum_{w' \in \Sigma} \exp(\mathbf{u}_{w'}^\top f_t([\mathbf{w}_T;\mathbf{z}]))}, & \text{for } r \in T \to w \end{cases} \tag{2}$$

where $[\cdot;\cdot]$ is vector concatenation. $f_t(\cdot)$ and $f_s(\cdot)$ are multi-layer perceptrons (MLPs) which are used to encode root and terminal rules.[8]

We have defined the form encoded syntactic knowledge will take as a C-PCFG, $\langle \mathcal{G}, \boldsymbol{\pi} \rangle$. We now define the procedure by which the model will *learn*, or induce, it. Here is our syntactic learning objective.

During grammar induction, our objective is to find the $\boldsymbol{\pi}$, or set of rule probabilities, that maximizes the likelihood of each sentence in our corpus, $s \in \mathcal{C}$. The likelihood of a sentence under our C-PCFG is the sum of the probability of every possible tree derivation t for s under our grammar $\mathcal{G}$ and conditioned on $\mathbf{z}$, or more formally:

$$p_\theta(s|\mathbf{z}) = \sum_{t \in \mathcal{T}_{\mathcal{G}}(s)} \prod_{r \in t_{\mathcal{R}}} \pi_{\mathbf{z}}(r) \tag{3}$$

where θ represents the parameters of our grammar model, $\mathcal{T}_{\mathcal{G}}(s)$ is the set of all derivations for s, and $t_{\mathcal{R}}$ the set of rules in a given tree t.

Since our learning objective is to maximize the likelihood of a each sentence irrespective of $\mathbf{z}$, we will need to marginalize over $\mathbf{z}$, taking the following integral:

$$p_\theta(s) = \int_{\mathbf{z}} \sum_{t \in \mathcal{T}_{\mathcal{G}}(s)} \prod_{r \in t_{\mathcal{R}}} \pi_{\mathbf{z}}(r) p_\gamma(\mathbf{z}) d\mathbf{z}$$

However, computing this integral is intractable. Instead, we can estimate a the log-likelihood of a sentence using variational Bayesian inference, by finding the maximum evidence lower bound (ELBO):

$$\log p_\theta(s) \geq \text{ELBO}(s;\phi,\theta,\gamma) = \mathbb{E}_{q_\phi(\mathbf{z}|s)}[\log p_\theta(s|\mathbf{z})] - \text{KL}[q_\phi(\mathbf{z}|s) \parallel p_\gamma(\mathbf{z})], \tag{4}$$

where $q_\phi(\mathbf{z}|s)$ is our variational approximation of the posterior parameterized by ϕ, in practice computed using another neural network.[9]

We can therefore fully define our syntactic objective as:

$$\mathcal{L}_{\text{syntax}}(\mathcal{C};\phi,\theta,\gamma) = -\sum_{s \in \mathcal{C}} \text{ELBO}(s;\phi,\theta,\gamma) \tag{5}$$

This objective will induce a C-PCFG, $\langle \mathcal{G}, \boldsymbol{\pi} \rangle$, by estimating the maximum log-likelihood of each sentence in our corpus $\mathcal{C}$ using the variational inference method. This approach amounts to turning the grammar induction problem into a parameter estimation problem, which we can effectively do via gradient descent using our neural network implementation.

In practice, we allow our grammar $\mathcal{G}$ to contain up to 30 non-terminal categories, 60 pre-terminals, and a vocabulary of 2000 terminals/words.[10] These numbers means that there are a total of up to 120,000 (60×2000) terminal rules, 243,000 ($30 \times (30 + 60) \times (30 + 60)$) non-terminal rules, and 30 root rules in our grammar. During grammar induction the model should quickly learn which of these rules are productive and thus more likely, narrowing down our grammar to a small subset of them in the end.

The semantic objective

We would like to represent semantic knowledge as a semantic embedding space, where visual information paired with syntactic forms can be encoded. To learn such a space, we take inspiration from previous work on vision-language model encoders (VLMs), such as CLIP (Radford et al., 2021). VLMs learn a joint text and image semantic embedding space by trying to maximize the similarity between text embeddings and image embeddings in sentence-image pairs (Kiros, Salakhutdinov, & Zemel, 2014). Unlike with VLMs, we do not want to simply take a text embedding for a complete sentence, but instead would like to have a way to represent the syntactic structure of a sentence in our embedding space as well. Thus, we now define the procedure for determining the similarity between an image and the predicted syntactic structure of a paired sentence.

First, we need a way to embed images. In previous work, this was done by taking the final layer of a pretrained vision model, such as ResNet-101, as an image embedding (Shi et al., 2019; Zhao & Titov, 2020). However, ResNet models like most vision models are trained in a supervised manner using class labels for images. In other words, these representations already encode categorical biases towards specific lexical terms. Given that our joint learning model serves as a cognitive model, we want to avoid introducing any initial linguistic biases into image representations. Thus, we train our own vision network on our dataset using an unsupervised learning algorithm called SimCLR (Chen, Kornblith, Norouzi, & Hinton, 2020) which requires no class labels at all.[11] We then extract an unbiased set of all image representations $\mathcal{V}$ from our custom vision model, here a custom pretrained self-supervised ResNet-50 model, $\texttt{net}_{\text{image}}$:

$$\mathcal{V} = \texttt{net}_{\text{image}}(\mathcal{I}),$$

where $\mathcal{I}$ are the images. These pretrained embeddings are fixed, so we do not want to directly apply our semantic objective to them. Instead,

[7] A reviewer asked whether such an assumption should be considered cognitively plausible and we think this is a fair question. Human languages are thought to be context-sensitive languages (Shieber, 1985), meaning shared contextual information is likely necessary to learn some of their rules. A fairer question may thus be whether complete context-freeness is cognitively plausible. The assumptions made with C-PCFGs find themselves somewhere in the middle: benefiting from the efficiency context-free rules offer for grammar induction over context-sensitive ones, but having some contextual information available in the form of a single shared whole-sentence representation.

[8] Following Kim et al. (2019), no such MLP is used with non-terminal rules. See Appendix "Model implementation" for detailed description of the MLP architectures.

[9] For the specifics of its implementation see Appendix "Model implementation".

[10] The number of categories is based on previous work (Kim et al., 2019; Zhao & Titov, 2020). We also tested other values (20,40) and did not find that it affected the results significantly. The vocabulary size was based on the number of unique words in our dataset, discounting typos. Other hyperparameters are available as the default settings in our code repository [Anonymized].

[11] SimCLR is a visual contrastive learning algorithm which works in the following way: given some target image, create a transformed version by randomly applying some set of image transformations such as crop, rotate, distort color, and/or add Gaussian noise. Then, over the course of training try to maximize the similarity of target and transformed images embedding pairs while minimizing the similarity with unmatched distractor images.

we would like to have representations which are mutable during the course of language learning. We thus make a distinction between our image representations and semantic representations, analogous to the distinction between visual perception, which is independent of language, and abstract representation and categorization of visual features, which may be influenced by language. We extract a semantic representation $\mathbf{m}$ from each image embedding $\mathbf{v} \in \mathcal{V}$ using our semantic encoder f_m, a trainable MLP:

$$\mathbf{m} = f_m(\mathbf{v}).$$

Second, for our semantic objective we must have a way to compare the similarity of our semantic image representations to the predicted syntactic structure of sentences. Following previous work, instead of using a text encoder for whole sentences, we encode each constituent in a sentence independently and compare them to our semantic representation $\mathbf{m}$ (Shi et al., 2019; Zhao & Titov, 2020). In practice, to ensure that our joint learning model is fully differentiable, we do this comparison with every possible word span in a sentence and then weigh each word span's relative importance by its marginal likelihood under the induced grammar – higher likelihood spans most likely being relevant constituents, while low probability ones most likely being irrelevant.

Thus, we define a semantic objective for a single sentence-image pair $(s, \mathbf{m})$ as follows:

$$\ell_{\text{pair}}(s, \mathbf{m}|\theta) = \sum_{c \in \text{spans}(s)} \text{match}(c, \mathbf{m}|\theta), \tag{6}$$

where $\mathbf{c}$ is a possible constituent, or word span, and $\text{spans}(\cdot)$ is a function which returns all possible spans of consecutive words from sentence s of length l where $1 < l < |s|$, and recall θ are the grammar model parameters. As for our $\text{match}(\cdot,\cdot)$ function, it is a weighted version of the contrastive loss traditionally used in training VLMs:

$$\text{match}(c, \mathbf{m}|\theta) = p_\theta(c|s, \mathbf{z}) h(\texttt{biLSTM}(c), \mathbf{m}), \tag{7}$$

where $p_\theta(c|s, \mathbf{z})$ is the marginal probability of a constituent, or its overall probability across all possible $t \in \mathcal{T}_\mathcal{G}(s)$,[12] which weighs the hinge loss function, $h(\texttt{biLSTM}(c), \mathbf{m})$. We encode each constituent independently using a single-layered `biLSTM` language model. The hinge loss – a traditional contrastive learning loss used with VLMs – then tries to maximize the similarity between matched constituent representations and semantic representations, while minimizing the similarity of unmatched pairs:

$$h(\mathbf{c}, \mathbf{m}) = \left[\text{sim}_{\cos}(\mathbf{c}', \mathbf{m}) - \text{sim}_{\cos}(\mathbf{c}, \mathbf{m}) + \epsilon\right]_+ + \left[\text{sim}_{\cos}(\mathbf{c}, \mathbf{m}') - \text{sim}_{\cos}(\mathbf{c}, \mathbf{m}) + \epsilon\right]_+, \tag{8}$$

where $\mathbf{c}$ is the constituent representation, $[\cdot]_+ = \max(0, \cdot)$, ϵ is a constant margin, and $\mathbf{c}', \mathbf{m}'$ are negative examples, or unmatched constituent and meaning representations from a different sentence-image pair. Intuitively, we want our model to learn to represent semantic knowledge in an semantic embedding space where matched constituent-image pairs are closer in space than unmatched ones.

The complete semantic objective is then simply the sum of the pairwise objective across all sentence-image pairs.

$$\mathcal{L}_{\text{semantics}}(C, \mathcal{V}; \theta) = \sum_j \ell_{\text{pair}}(s^{(j)}, f_m(\mathbf{v}^{(j)}); \theta) \tag{9}$$

The complete model and ablations

Now that we have defined both how syntactic and semantic knowledge is represented and how we can learn these representations via our syntax and semantics objectives, we can bring all these components together in our joint-learning model. We will additionally consider some ablated versions of the model to better understand the role of each of these components over the course of joint learning.

[12] To see how we can derive this value from our derivation tree distribution $p_\theta(t|s, \mathbf{z})$ over trees for s, see Zhao and Titov (2020).

Joint-learning model: Simply put, the model optimizes both the syntactic objective and the semantic objective at the same time during training, giving us the joint loss:

$$\mathcal{L}_{\text{joint}}(C, \mathcal{V}; \phi, \theta, \gamma) = \alpha_1 \cdot \mathcal{L}_{\text{syntax}}(C; \phi, \theta, \gamma) + \alpha_2 \cdot \mathcal{L}_{\text{semantics}}(C, \mathcal{V}; \theta), \tag{10}$$

where α_1, α_2 are constants, here both equal to 1. Since the syntactic and semantic objectives are interdependent, during learning they will affect the joint model's learning trajectory by mutually constraining updates to the grammar and the semantic embedding space. On the one hand, the semantic objective will push the grammar model to favor rules which derive trees containing constituents that can be more easily visually represented. This is because the semantic objective maximizes the similarity between representations of word spans and images weighted by the marginal probability of word spans. Thus, there are two ways to increase it: (1) updating the parameter values of matched word span and image embeddings to be closer in semantic space and/or (2) increasing the weight or marginal likelihood of word span embeddings that are already similar to their semantic counterparts. On the other hand, the syntactic objective will determine the distribution over constituents being mapped to semantic space. Again, this is because the semantic objective is weighted by the marginal probability of word spans under the grammar. The grammar is updated as a function of the syntactic objective, which in turn determines the marginal likelihood of word spans for the semantic objective. We illustrate the model and the interconnected relation between objectives in Fig. 3.

In using our model as a cognitive model, we are making the following assumptions about the acquisition of syntactic and semantic knowledge: (1) we start with the prior knowledge that there are such things as syntactic categories (though what they represent and how many there are is unknown); (2) grammatical structure can be represented by a PCFG; (3) meaning is grounded in visual representations. The first assumption is feasible under most linguistic theories and theories of acquisition. The second and third, are likely simplifications of children's learning environment: PCFGs may not have enough explanatory power for natural language (Huybregts, 1984; Shieber, 1985), and children ground language in embodied experiences going way beyond still images. By definition, a model is a simplified exemplification of a process. We acknowledge that our model does not capture all the complexities present in a child's naturalistic learning environment. However, we are using our model to study specific research questions and not language learning as a whole. We believe that this joint-learning model can demonstrate how joint learning objectives for language can aid language acquisition overall and explain bootstrapping phenomena, like syntactic and semantic bootstrapping.

Semantics-first model: Our first ablated model starts by optimizing the semantic objective only. Since this model does not initially have access to syntax or predicted structures, it uses the similarity between whole sentences and images, as opposed to the similarity between the constituents and an image.[13] In other words, it uses the traditional image-caption matching loss used in VLM models like CLIP (Radford et al., 2021), having access to only the order in which words appeared but no structure beyond that. Halfway through training, we add the syntactic objective, resulting in the regular joint-learning objective. This first ablation will help us understand the impact of having access or not to syntactic representations from the start of learning, showing us what happens when semantic knowledge is initially acquired independently of syntactic knowledge.

[13] We also tried a version that uses the semantic loss defined above but found that it made no difference in model performance, though it required much more computation. This is because the C-PCFG is initialized uniformly, meaning all rules have equal likelihood and thus all possible trees or spans of equal length are equally likely. Since the C-PCFG does not update for the first half of training in the semantics-first model, staying uniformly distributed, nothing is gained from applying the loss over spans as opposed to the whole sentence.

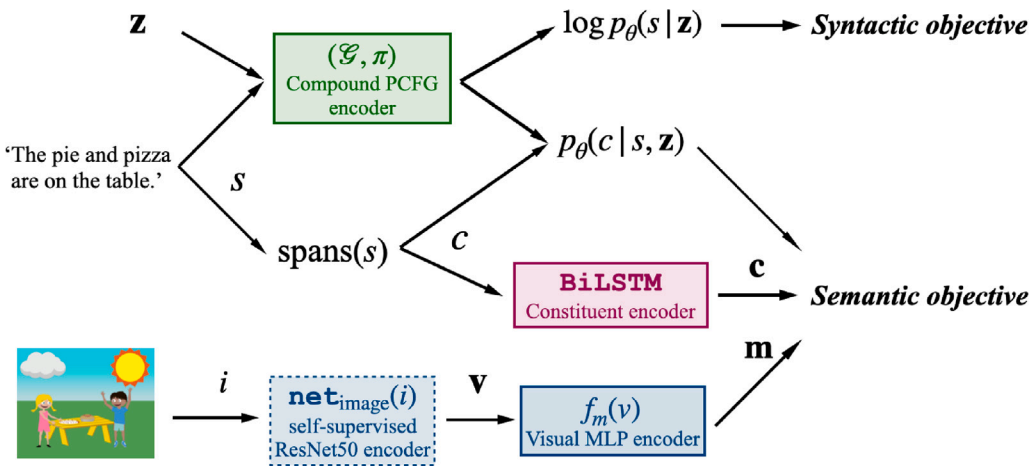


Fig. 3. The joint model architecture.

Syntax-first model: Our second ablated model is initially trained solely on the syntactic objective, or without visual grounding, in other words using the original C-PCFG grammar induction objective from Kim et al. (2019). Halfway through training, we add the semantic objective, resulting once again in the regular joint-learning objective. This second ablation allows us to determine the impact of having access or not to semantic representations from the start of learning, showing us what happens when syntactic knowledge is initially acquired independently of semantic knowledge.

By comparing the joint-learning model to the syntax-first and semantics-first ones, we hope to demonstrate the importance of an interdependence between syntactic and semantic representation learning, supporting our proposal for a unified view of bootstrapping theories. Finally, we compare the joint-learning model to a model with oracle knowledge of image content, which we call the visual-labels model.

Visual-labels model: This model is trained using the same objective as the joint-learning model, however, instead of using image representations from our unsupervised pretrained image vision model, it uses label vectors which we extracted using the Abstract Scenes metadata. These label vectors encode the coordinates and rotation of all objects and agents in images, as well as the physical position and facial expression presented by agents. Our unsupervised pretrained image encodings may not have learnt to extract all of the information contained in the label vectors. Thus, the visual-labels model can give us a sense for how important the clear perception of referents is for our evaluation tasks.

Experiment 1: Semantic bootstrapping and joint learning

This experiment answers our first research question: can access to visual-grounding and the ability to learn semantic representations in a joint learning setting facilitate learning grammars that generalize better to unseen contexts? In other words, akin to semantic bootstrapping, does having access to visual-grounding, or the ability to map visual meanings to strings, help models learn syntactic categories and productive syntactic rules over language. We compare four models, the joint-learning model, the semantics-first model, the syntax-first model, and the visual-labels model. For each model we train five versions from different random seeds. Based on Pinker (1984) proposal, access to semantic representations is the basis for learning meaningful syntactic categories. We thus expect models with access to semantic representations from the start to learn more meaningful grammars over language, with syntactic categories that better mirror those described in language acquisition and linguistic theories. Furthermore, we hypothesize that joint learning helps constrain the hypothesis space considered over grammars leading to less variation across model runs. We thus expect the joint-learning and the semantics-first models to learn better grammars than the syntax-first model and that the joint-learning model should show less variation across runs than the semantics-first or syntax-first models.

Evaluations

To determine what makes an induced grammar a *better* grammar, we report two evaluations. The first follows previous work in grammar induction and compares the internal branching structure, excluding the root and leaf branches (which are deterministic), of the most likely parse for each sentence under the induced grammars to gold standard parses. We also compare our models to fully left-branching and fully right-branching baselines. The second evaluation compares the induced pre-terminal categories of models to gold syntactic categories to determine whether the induced categories over words correspond to syntactic categories traditionally used in linguistics. This second evaluation is inspired by semantic bootstrapping theory, which was intended to explain both how children learn syntactic primitives, especially formal categories over words, and what distributional information they must attend to in their input to do so Pinker (1984). We can compare the impact of visual-grounding and objective functions on the models ability to induce meaningful syntactic categories, following Pinker's proposal for a model of direct evidence of the semantic bootstrapping hypothesis.

The gold parses, or rather in this case silver parses, were automatically extracted using the Berkeley Neural Parser (Benepar) (Kitaev, Cao, & Klein, 2019; Kitaev & Klein, 2018), which derives constituency parses for sentences and uses part-of-speech tags from the SpaCy library (spaCy, 2020) as lexical categories. This approach is also taken in Shi et al. (2019) and Zhao and Titov (2020). To create a correspondence between the part-of-speech tags on these parse trees and syntactic categories, we used a custom mapping provided in Appendix "Syntactic category mappings". We train five different random seed runs for each model.

Results

Evaluating syntactic structure

We compare the induced trees on the in-distribution test sentences to gold parses, as well as to fully left-branching and right-branching trees. We use a standard metric for this comparison: mean span F1 score across trees. The span F1 score of a predicted tree is calculated by considering all the intermediate constituents of more that one word in a sentence as determined by its branches. Fig. 4 demonstrates this process. We report final scores in Table 1. The model which induces trees closest to the gold parses is the joint-learning model. Given

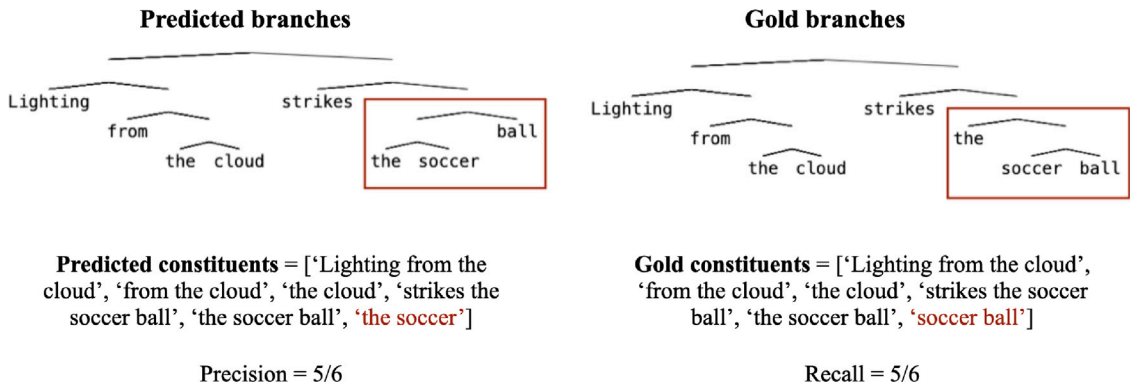


Fig. 4. Span F1 score is the harmonic mean of constituent precision and recall between a predicted tree and its gold counterpart. In this example the span F1 score is 0.83.

Table 1
Mean span F1 scores by model over test sentence trees for all random seed runs combined. Standard deviations across test trees in parentheses. T-tests between the joint-learning model scores and all other models indicate that increase in performance is significant in all cases (p = [>0.001]).

Model	Gold parses
Right-branching	0.85 (0.18)
Left-branching	0.08 (0.12)
Joint-learning	**0.90 (0.16)*****
Semantics-first	0.75 (0.24)
Syntax-first	0.42 (0.21)
Visual-labels	0.87 (0.17)

that these are English sentences, the gold parses are quite similar to right branching trees, with on average 85% of internal branch span correspondence. These scores are taken after 30 epochs of learning, though in Fig. 5, we additionally plot the changes in mean span F1 scores between predicted trees and gold parses throughout learning for each model. The vertical dashed line in this figure marks the halfway point where syntax-first and semantics-first models switch to a joint-learning objective (adding in either the semantic loss or the syntactic loss respectively). Joint learning clearly leads models to successfully induce grammars over language. Whether joint learning is paired with the self-supervised image embedding or with the gold label image embeddings (visual-labels model), does not seem to matter. Interestingly, with the semantics-first model, which is simply at chance for the first half of learning since it does not yet have access to the syntax loss for learning grammar, we see that once we introduce the joint loss, it is able to learn reasonable grammars, however there is much more variation across runs. This observation supports the hypothesis that joint learning is in part most successful because it can lead to mutual constraining of the hypothesis space for both syntactic and semantic learning. As for the syntax-first model, without the availability of visual grounding initially, it induces grammars that are very different from the one used for the gold parses.[14] Even once visual grounding is made available at the halfway point, the model can no longer recover, having limited itself to a certain hypothesis space over categories and rules that is too far from those used in deriving the gold parses.

In Fig. 6, we compare example induced trees from the joint-learning model to the respective gold parses to see where differences lie. One of the main differences is in noun phrases with adjectives, where we can clearly see that predicted trees have determiners subordinate to nouns, while gold parses have nouns subordinate to determiners. This difference is reminiscent of the noun phrase (NP) versus determiner phrase (DP) debates in theoretical linguistics (Köylü, 2021). Another difference seems to be in the treatment of coordinate structures in subject position, where predicted trees have the first conjunct selecting for the sentence containing the second conjunct as subject, instead of having the conjunction as a whole serving as subject.

Comparing induced syntactic categories

The first evaluation considered the quality of the induced grammar. In this second evaluation, we look at the pre-terminal or lexical categories having been learnt over words. We compared the induced categories for words in predicted parses, to the syntactic categories associated with words in the gold parses. Since each random seed run may learn a different mapping between categories, we first consider each run separately and visualize how well predicted categories map to gold ones using a normalized contingency table. In Fig. 7, we plot the proportion of words in each syntactic category that was mapped to a given pre-terminal category — of which there were up to 60.[15]

With the exception of the syntax-first model, others seem to distribute the majority of words into 10 to 20 pre-terminal categories. In the case of the joint model, most syntactic categories map to distinct sets of pre-terminals, suggesting that the model was able to learn syntactically meaningful lexical categories over words. There is some overlap between nouns, proper nouns, and pronouns, though since these types of words tend to occupy similar syntactic positions, it is not too surprising. The semantics-first, syntax-first, and visual-labels models were also able to induce meaningful lexical categories. We note that the syntax-first model seems not to have any clear correspondence in its pre-terminals to verbs. Since verbs are so important to sentential structure, the lack of a clear verb category or set of categories may in part explain why this model also fails to induce grammars which resemble our gold parse grammar in the first evaluation.

Additionally, we measure the quality of models' induced lexical categories using V-measure, an entropy-based cluster evaluation metric (Rosenberg & Hirschberg, 2007). Similarly to F-scores, V-measure is a weighted harmonic mean of two cluster quality metrics, homogeneity and completeness.[16] Intuitively, homogeneity measures how well induced lexical categories map to labeled syntactic categories (i.e. do all words in C52 map to a single syntactic category, like *modal*?), while completeness measures how well labeled syntactic categories map to induced lexical categories (i.e. do all words labeled *adjective* map to a single pre-terminal category, say C35?). Given that we are interested in inducing homogeneous lexical categories where words all share particular syntactic properties, this measure matters most to us. Completeness

[14] Though these results may seem low, they are in line with previous findings on other corpora, where non-visually grounded C-PCFG mean F1 branching scores range between 0.36 and 0.55 (Kim et al., 2019; Zhao & Titov, 2020).

[15] The same plots for other random seeds as well as additional analyses using mean Jensen–Shannon divergence between predicted and annotated category distributions are available in Appendix "Syntactic category mappings".

[16] All these measures are bounded between 0 and 1, where in general higher values are considered better.

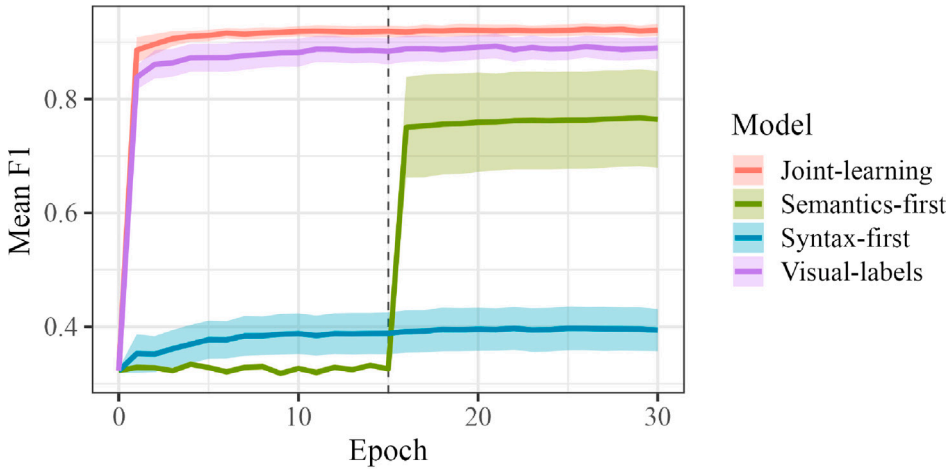


Fig. 5. Mean Span F1 scores on test sentences by model during learning. Shading represents standard error across 5 runs. Dashed line represents point in time where semantics-first and syntax-first models switch to joint-learning loss function.

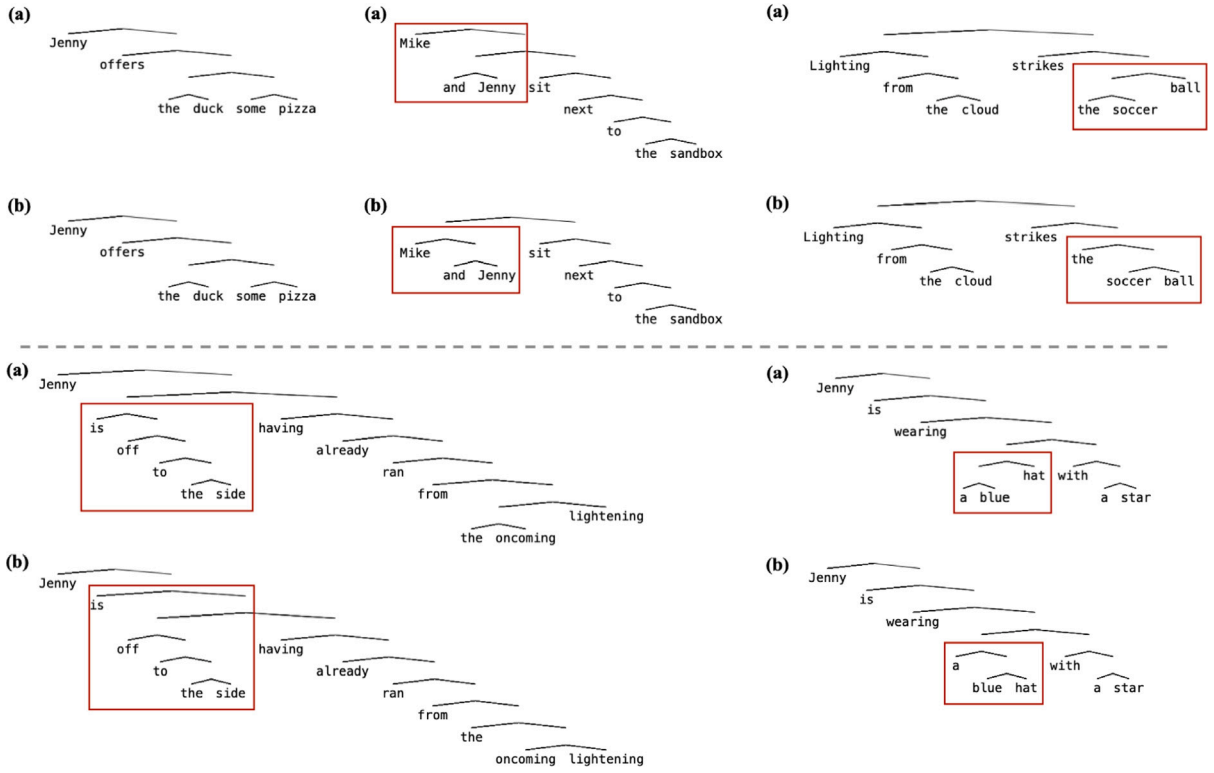


Fig. 6. Examples of (a) induced trees from the joint-learning model and (b) gold parses from the Berkeley Neural Parser. Red boxes highlight discrepancies between predicted and gold trees.

may matter less here since the level of precision of our labeled syntactic categories is arbitrary and models may in fact be inducing more precise categories — for example, had we chosen to separate the category *verb* into more subcategory labels such as transitives, intransitives, and ditransitives, it is possible that completeness would then be higher. We report mean V-measure, homogeneity, and completeness across model runs in Table 2.

The results in Table 2 confirm that all models were able to learn meaningful lexical categories and that joint learning can lead to successful syntactic category learning. We note that the syntax-first model had the lowest homogeneity measure, while the semantics-first and

Table 2
Mean V-measure cluster evaluation results for predicted pre-terminal categories and labeled syntactic categories across random seed runs. V-measure is a β-weighted harmonic mean of homogeneity and completeness. Here $\beta = 0.3$ to weight homogeneity more importantly. Standard deviations are in parentheses.

Model	V-measure	Homogeneity	Completeness
Joint-learning	0.82 (0.01)	0.89 (0.01)	0.44 (0.02)
Semantics-first	0.85 (0.01)	0.90 (0.01)	0.49 (0.03)
Syntax-first	0.82 (0.01)	0.87 (0.01)	0.50 (0.03)
Visual-labels	0.83 (0.02)	0.90 (0.01)	0.46 (0.04)

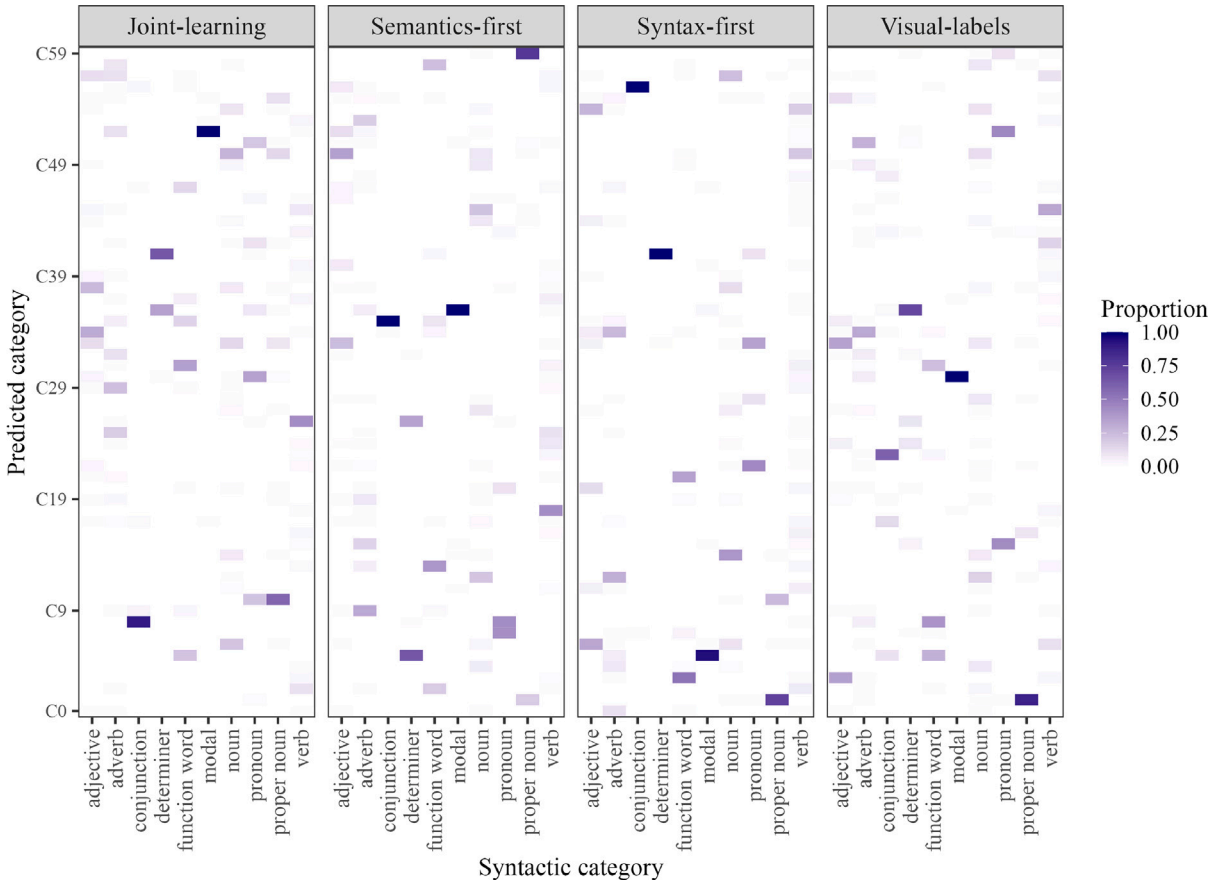


Fig. 7. Proportion of predicted category to syntactic category mappings on all sentences. Models with random seed 1018 (other random seed results available in supplementary materials).

visual-labels models had the highest.[17] This observation further suggests that access to visual grounding early in learning as a proxy for semantic meaning can improve a model's ability to learn syntactically relevant lexical categories, as predicted by semantic bootstrapping theory.

Experiment 2: Syntactic bootstrapping and joint learning

In this experiment we answer our second research question: can access to linguistic structure and the ability to learn grammar in a joint learning setting facilitate learning and interpreting novel words and contexts? Akin to syntactic bootstrapping, we would like to test if syntactic structure can help models interpret novel word meanings in context. Like in our first experiment, we compare four models: the joint-learning model, the semantics-first model, the syntax-first model, and the visual-labels model, training five versions from different random seeds. This experiment is inspired by experimental designs from studies on syntactic bootstrapping (Fisher et al., 2020). We hypothesize that joint learning and access to syntactic structure should better a model's ability to interpret novel words and contexts. Therefore, we expect joint-learning models – both the visual-labels model and joint-learning model with self-supervised image embeddings – to perform best, while the semantics-first and syntax-first models should see an increase in performance after introducing joint learning.

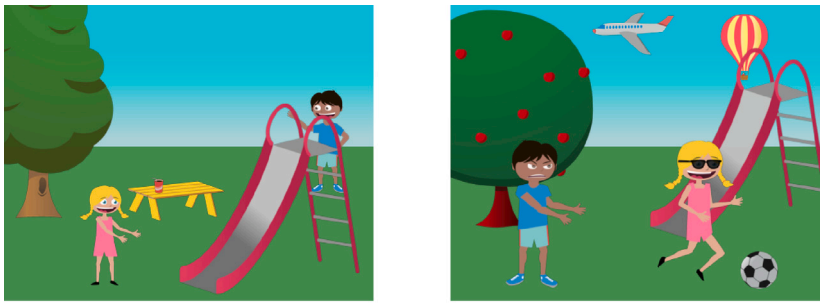
[17] A t-test between the semantics-first and syntax-first confirms that this difference is significant ($t([4]) = [4.58]$, $p = [>0.05]$), though not for the visual-labels model ($t([4]) = [2.46]$, $p = [0.06]$).

Evaluations

There are two evaluations for this experiment: The first involves matching a novel sentence containing a never before seen verb to visual scenes (see Fig. 8), the second involves matching a visual scene to sentences which are minimally different, only interchanging semantic roles (see Fig. 9).

The first evaluation is an out-of-distribution test done with models for which we have withheld all test sentences containing the 30 different verb stems (10 transitive-animate, 10 transitive-inanimate, 10 intransitive) from their training data. The sentences containing these held out stems served as test sentences to evaluate whether or not models could interpret novel verbs and contexts by correctly identifying the corresponding images from a pair of images. The test examples are constructed by pairing one transitive sentence with an intransitive one, where their respective images then serve as each other's target and distractor images. This test is based on the previously mentioned nonce verb learning experimental paradigm shown in Fig. 1. We note that our evaluation does differ in some ways. Since our test examples are built using existing images and sentences from the Abstract scenes dataset, they are not necessarily minimal pairs, so it is possible at times to use additional sentential cues to help correctly identify the target image. For example, in Fig. 8, models could use the presence of the adverb 'happily' to help pick the correct image, since in the distractor image, Mike does not seem happy. Our evaluation setup however allows us to have many test items, 1436 to be exact. We cannot fully control for additional factors beyond the syntactic structure surrounding novel verbs contributing to models correctly matching sentences to images. For this reason, we additionally introduce a second evaluation.

Syntactic bootstrapping theory argues that children can use their understanding of argument structure and semantic roles to help them



'Mike is happily climbing the pink slide'

Fig. 8. First evaluation: matching sentences with novel verbs to images. Example test item from transitive-inanimate condition. Models have never seen the verb 'to climb' during learning.

(a) 'Jenny is waving to Mike'
(b) 'Mike is waving to Jenny'

Fig. 9. Second evaluation: matching semantic roles to images. Example test item with semantic role alternation. This is an in-distribution evaluation which tests whether models can distinguish semantic roles using sentence minimal pairs.

interpret and learn the meanings of novel verbs (Gleitman, 1990). To determine if models have learnt to distinguish semantic roles and thus whether they could be using a similar strategy to correctly interpret novel verbs in the first evaluation, we use a second follow up evaluation taken from Nikolaus and Fourtassi (2021a), illustrated in Fig. 9. Here, models are given an image with a transitive action and two minimally different sentences, only differing in that the agent and patient roles have been reversed. Models must then correctly identify the sentence which contains the appropriate semantic role assignment based on the image. This second evaluation is an in-distribution test using carefully constructed sentence minimal pairs with known verbs that were not in the original Abstract Scenes dataset. It is more controlled and can pinpoint how much models understand about the argument structure of known verbs and their semantic roles, however it has only 50 test items.

Results

Matching sentences with novel verbs to images

This first evaluation measures how well models can interpret sentences containing novel verbs and whether they can distinguish between transitive and intransitive actions. To determine a model's preference, we first extract a *tree embedding*, or a representation of the test sentence structure under the model, by taking the average of all span embeddings normalized by their likelihood under the induced grammar (Eq. (11)).[18]

$$\mathbf{t}_s = \frac{\sum_{c \in \text{spans}(s)} p_\theta(c|s, \mathbf{z}) \texttt{biLSTM}(c)}{|\text{spans}(s)|} \tag{11}$$

We then return the cosine similarity between our tree representation and the semantic representations of both the target and distractor images, taking the one with the highest similarity score as the model's choice (Eq. (12)).

$$\hat{y}_{verb} = \max(\text{sim}_{\cos}(\mathbf{t}_s, \mathbf{m}_{\textbf{target}}), \text{sim}_{\cos}(\mathbf{t}_s, \mathbf{m}_{\textbf{distractor}})) \tag{12}$$

We report the average proportion of correctly identified target images, or the mean matching score, across 5 random seed models and 1436 test items. We plot models' matching scores throughout training in Fig. 10. Chance performance is 0.5 since this is a balanced binary choice task.

All models successfully learn to map novel verb sentences to their respective images the majority of the time. We note that the visual-labels model which has gold labels as visual encodings for image content does much better than other models. All the other models as previously described use self-supervised visual encodings which may not have learnt to encode all the necessary visual information. Still, we see that the joint-model's performance consistently increases while both the syntax-first (at chance initially since it has no access to visual encodings) and the semantics-first models see a jump in performance after the introduction of the joint-learning loss at the half way point.[19] Performing a paired t-test for the semantics-first model right before and after having introduced joint learning confirms that this jump is significant ($t([4]) = [5.04]$, $p = [>0.01]$).

We additionally plot model performance with respect to target novel verbs being either transitive or intransitive in Fig. 11, each with 718

[18] In the case of the semantics first model, for the first half of training, we simply use the complete sentence embedding $\texttt{biLSTM}(s)$ as the test sentence representation to speed up evaluation time. We also tried using the tree representation, but found it made no difference for results since in the semantics first model, the grammar is initially uniformly distributed – meaning that all rules are equally likely, and thus all of trees for a sentence are equally likely as well – having therefore no effect on the relative ranking of the images to sentence similarity scores.

[19] Astute readers will have noticed that though the syntax-first model saw no improvement after the introduction of joint learning in experiment 1 on semantic bootstrapping in Section "Experiment 1: Semantic bootstrapping and joint learning", here the syntax-first model does catch up after the introduction of joint learning. This observation suggests that is not so much the particular verb phrase structure we learn for transitive and intransitive verbs that matters for syntactic bootstrapping (e.g. "[Agent [verb [Patient]]]" and "[Agent [verb]]"), but simply that they be distinguishable, for example "[[[Agent] verb] [Patient]]" versus "[[Agent] verb]", may not correspond to our gold parses, but they still make a distinction between verb type structures and may have relevant constituents to map to semantic representations.

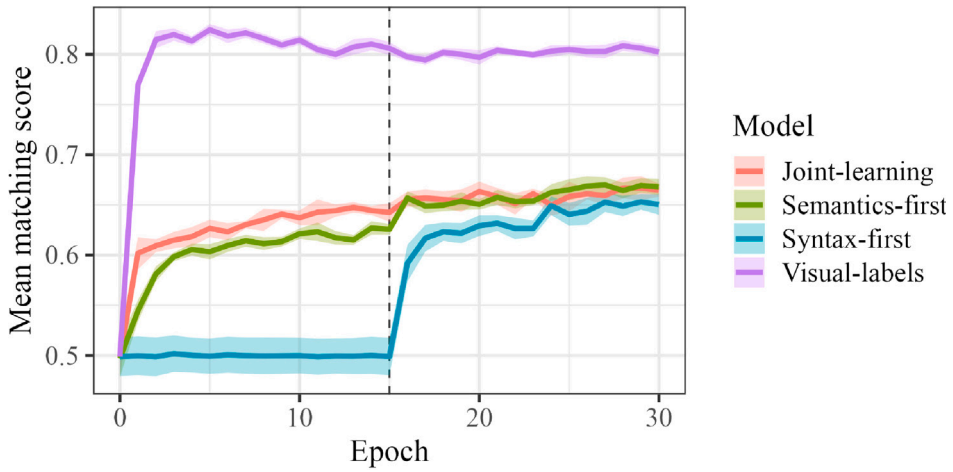


Fig. 10. Matching novel verbs: Mean matching scores on sentences with out-of-distribution verbs by model during learning. Shading represents standard error across 5 runs.

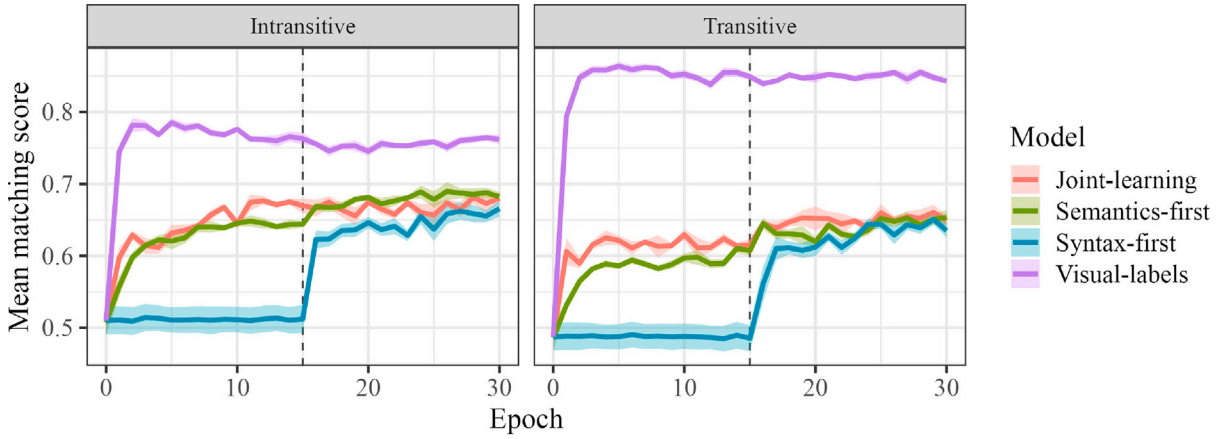


Fig. 11. Matching novel verbs: Mean matching scores on out-of-distribution sentences by verb type and by model during learning. Shading represents standard error across 5 runs.

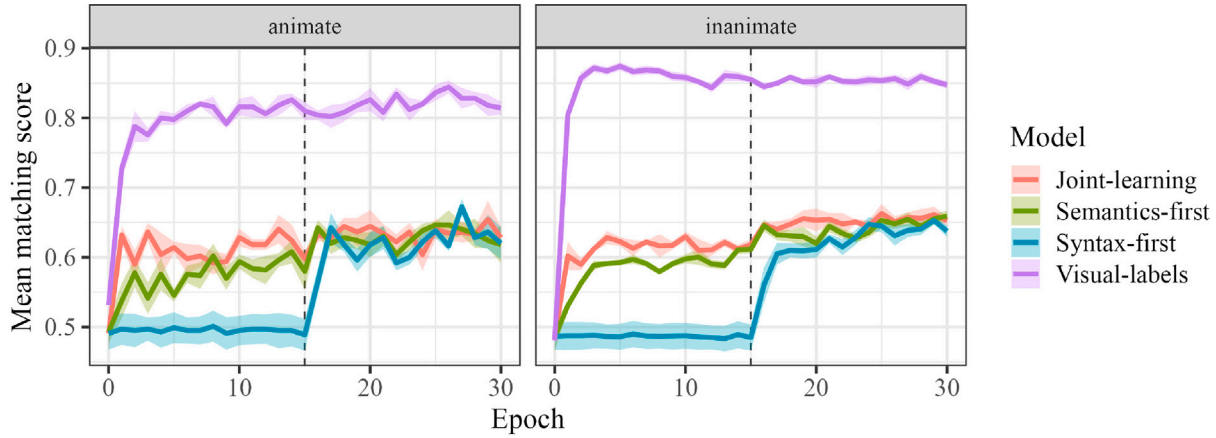


Fig. 12. Matching novel verbs: Mean matching scores on out-of-distribution transitive sentences by object type and by model during learning. Shading represents standard error across 5 runs.

test items. We see no meaningful difference between conditions for the joint-learning model and syntax-first model. The jump in performance observed in the previous figure for semantic-first models seems to be in the transitive condition more prominently. Transitive sentences contain more structure and require more mastery of semantic roles to properly interpret them; access to syntactic information via joint learning may be especially important in this condition, showing evidence of model behavior that is consistent with the syntactic bootstrapping hypothesis. Interestingly, the visual-labels model struggles more in the intransitive condition than the transitive one. Since this difference is not present in the other models, this observation likely indicates that the visual-labels model is relying more heavily on additional visual information in some

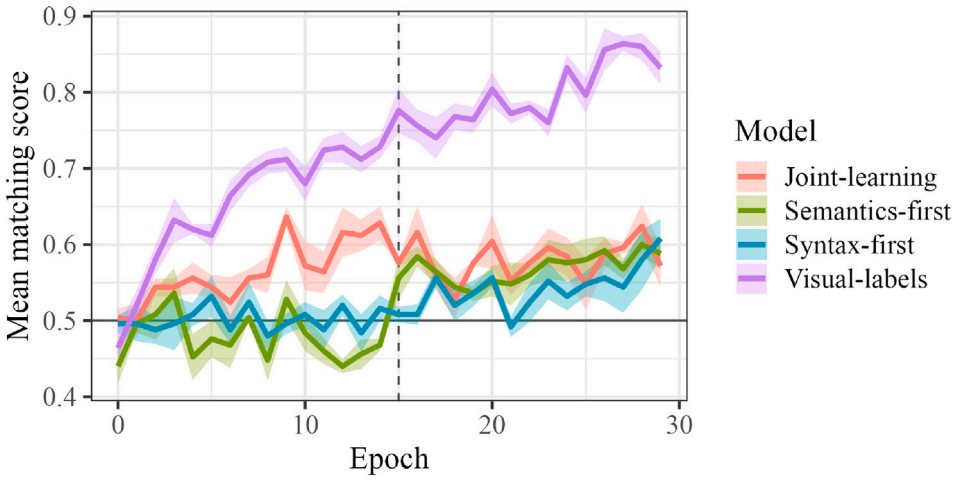


Fig. 13. Matching semantic roles: Mean matching scores on semantic role test sentences by model during learning. Shading represents standard error across 5 runs.

contexts over syntactic information. Since transitive sentences and their corresponding images are likely to contain more distinct referents, especially in the transitive-inanimate object condition, the difference in performance between these two conditions could be explained by the model relying on distinctive visual cues (eg. the presence of a soccer ball in one image versus another). The same does not seem to be true of other models.

Within the transitive condition, we compare performance between trials that had an animate (person or animal) or inanimate object in Fig. 12. We find more evidence in favor of our hypothesis about the visual-labels model, as it performs a little better in the inanimate condition which contains more distinctive referents. While other models once again show little difference, learning to correctly respond the majority of the time regardless of object type. The curves for transitive animate test items are slightly noisier because there are fewer examples of these in the test set, 99 over 618 for the inanimate object examples.[20]

Matching semantic roles to images

To determine if models are sensitive to semantic roles, we measure their ability to distinguish between minimally different sentences containing reversed semantic roles given an image. Like with the previous evaluation, we first extract tree representations for both the target and distractor sentences following Eq. (11). We then compare their similarity scores with the test image using Eq. (13).

$$\hat{y}_{role} = \max(\text{sim}_{\cos}(\mathbf{t_{target}}, \mathbf{m}), \text{sim}_{\cos}(\mathbf{t_{distractor}}, \mathbf{m})) \quad (13)$$

We report the average proportion of correct matches across random seed runs and 50 test items, or models' mean matching score. Chance performance is once again 0.5. Fig. 13 shows models' performance throughout learning. Neither the syntax-first and semantics first models do much better than chance at first. This pattern is expected for the syntax-first model, but for the semantics-first model, it highlights the importance of syntactic structure for learning semantic roles. Once joint learning is introduced at the half-way point their performance starts to increase. The joint-learning model and visual-labels model see a consistent rise in performance, which plateaus around the halfway point for the joint-learning model, likely due to the limitations of its visual encodings.

Given the limited number of test examples, the performance curves are quite noisy. It is clear that this task is difficult and that all models but the visual-labels model are limited by the accuracy of their visual embeddings. Still, joint learning allows models to successfully learn to identify semantic roles in a many cases, and when not limited by visual perception, in most cases.

Discussion

We set out to show via a computational cognitive model that both semantic bootstrapping and syntactic bootstrapping effects arise as a result of the interplay between syntactic and semantic knowledge acquisition during joint learning. Semantic bootstrapping and syntactic bootstrapping are not necessarily meant to be independent of one another and in fact, as our experiments suggest, the strongest effects of syntactic and semantic bootstrapping arise when we view language learning as a joint inference problem, where both semantics and syntax are learnt simultaneously.

Akin to semantic bootstrapping, existing works on neural visually/semantically grounded grammar induction (Jin & Schuler, 2020; Li et al., 2024; Shi et al., 2019; Wan et al., 2022; Zhao & Titov, 2020) have found that access to images or LLM semantic embeddings can lead to moderate improvements in grammar induction; in this work, we found that our visually-grounded joint-learning model saw large improvements, learning much better grammars than a syntax-first or syntax-only model. Our large improvements over moderate ones seen in previous work may be due to the following differences. To start, other studies used image-caption datasets (e.g. MS-COCO; Chen et al., 2015; Lin et al., 2014) where captions are not complete sentences with main verbs for the most part, which may be important when considering grammar acquisition as a whole. Furthermore, our model not only learnt the grammar from scratch, but also its visual and semantic representations, leading to possibly better suited image representations for semantic bootstrapping like effects to occur. A final difference and possible limitation to our study is that we used synthetic images while previous studies used real world images. It is possible that synthetic images made the task of identifying visual feature easier. Children's input is undoubtedly noisier and richer than the data our model was exposed to, but image-caption datasets are no closer to their learning experience either.[21] Given our primary goal: to demonstrate how the interplay between syntactic and semantic acquisition can follow

[20] Transitive verbs taking inanimate objects are generally much more frequent in the corpus.

[21] As one of our reviewers astutely noted, children's input being noisier may also mean that in their particular learning contexts joint-learning may not be possible without stronger syntactic and semantic prior biases. This

from joint learning, using simulated children's book data presented a sufficient environment to examine these dynamics.

Akin to syntactic bootstrapping, we found that access to grammatical knowledge over the course of learning increased models' ability to interpret novel sentences with never before seen verbs as well as their ability to recognize different semantic roles. These syntactic bootstrapping effects were however not as strong as those observed for semantic bootstrapping. We found that the quality of our visual representations for images were in part to blame for this discrepancy — using informative visual labels instead of training visual representations from scratch significantly helped models interpret novel verbs and learn semantic roles. In future extensions of this work we hope to address this limitation by considering other methods for learning visual representations that may better emulate the visual features that are salient to children.

Joint learning from the start works because it helps mutually constrain related hypotheses spaces, here grammar and semantic representations. These types of constraints are likely necessary for human learners who are limited in terms of memory and processing capacity as well as amount of input evidence. Even with these limitations though, we learn language and better yet, we learn representations which allow to generalize and use language in completely novel contexts. The reason for our learning efficiency and generalization abilities may lie in our effective learning strategies, which we argue are built on joint learning.

Empirical evidence suggests that semantic and syntactic processing during language comprehension or production are not separable into distinct areas of the brain, but instead represent distributed processes which overlap across a wide region referred to as the language network (Fedorenko, Blank, Siegelman, & Mineroff, 2020; Fedorenko, Ivanova, & Regev, 2024; Hu et al., 2022; Shain et al., 2024). Furthermore, children's lexicon and their syntactic production abilities grow side by side during language development (Bates et al., 1994; Brinchmann, Braeken, & Lyster, 2019; Frank, Braginsky, Marchman, & Yurovsky, 2021). These results all support our proposal: that language learning is joint learning across many levels of linguistic representation. The acquisition of morphemes, words, syntax, semantics, pragmatics have for the most part been considered in isolation. However, if language learning is indeed a joint inference problem across many levels of linguistic structure, then future research in the field should try to understand *how* learning biases or constraints within these different levels arise as a function of joint learning. For example, how does the acquisition of semantic knowledge affect the acquisition of syntax? or how does learning morpheme boundaries interplay with the acquisition of semantic knowledge? Understanding how these constraints arise and interact, we suggest should be the next key direction in language learning debates.

Computational modeling is not new to the fields of language development and cognitive science. However, the ability to design models like the one in this paper as well as the access we now have to multimodal data—sound, image, video, text— may allow us to revisit with new perspective many of the research questions we still have about language learning. We have proposed that a promising way to do so is to think of language learning as a holistic problem involving joint inference over many different levels of abstract linguistic representation. We have shown how joint learning does not necessarily make learning *harder*, but can make it *easier* by mutually constraining the hypotheses being considered by a learner, helping them acquire the complex systems that are human languages. We hope that this work may convince other researchers in both cognitive science and AI that an important new direction for language modeling and learning research lies in considering the dynamics of joint inference over many input sources and modalities as well as levels of representation.

observation highlights that any conclusions made using a model are bound by the simplifying assumptions made in designing the model and experiments (as with any modeling work).

CRediT authorship contribution statement

Eva Portelance: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Siva Reddy:** Writing – review & editing, Resources, Funding acquisition. **Timothy J. O'Donnell:** Writing – review & editing, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eva Portelance reports a relationship with Microsoft Research that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Verb stem lists for data split

These are the stems of verbs which appeared at least 5 times in the corpus. We held out for testing all instances of the following verb stems:

Held out transitive verb stems *taking animate objects* – push, rescu, teas, argu, hug, warn, feed, meet, fight, invit – *taking inanimate objects* – drop, open, ride, pour, brought, prepar, toss, use, climb, rais (hand).

Held out intransitive verbs walk, hide, smile, cheer, laugh, slid, cri, danc, fell, crawl

The training data included verbs with the following stems:

Training verb stems is, wear, are, sit, hold, has, stand, play, want, fli, slide, hot, kick, run, sad, scare, swing, mad, rain, wave, picnic, grill, bat, see, catch, watch, go, throw, jump, afraid, look, was, tri, surpris, eat, set, will, threw, drink, upset, have, get, excit, like, come, hand, shine, worri, doe, be, hit, chase, cook, made, camp, had, talk, fall, wait, give, start, think, sat, float, put, got, hurt, help, took, wore, saw, growl, flew, were, call, love, lost, went, shock, carri, offer, make, take, roar, pet, toy, found, stole, came, let, land, enjoy, can, tell, yell, pretend, reach, slither, strike, startl, burn, say, would, fetch, shin, stuck, know, ran, caught, did, goe, move, stare, find, follow, could, share, do, bring, might, rest, show, leav, round, notic, built, feel, waiv, ruin, perch, grow, pick, frighten, stop, miss, bake, struck, warm, seem, scream, leg, cover, lay, thunder, snif, keep, stay, should, color, ask, fallen, cross, bite, face, held, said, done, blast, roll, pass, ate, bounc, taken, ripe, hate, gone, thrown, closer, gave, place, stood, seen, tie, care, wish, point, hope, begin, pitch, forgot, holdng, waddl, snuck, annoy, tire, steal, dig, barbecu, dri, wonder, shout, been, decid, soar, skip, frown, understand, glad, dress, approach, sneak, shade, lose, copi, alarm, build, finish, astonish, standng, block, touch, knock, amus, plan, confus, better, may, hover, stripe, jog, told, bore, need, head, listen, thrill, join, began, attack, trade, smell, grab, trip, frustrat, stit, storm, stolen, hidden, sleep, sail, snake, hear, kneel, launch, march, juggl, serv, protect, site, tast.

Appendix B. Model implementation

All models were trained on A100 Multi-Instance GPU partitions, using at most 32Gb of GPU memory. Each model took about 18 h to train and evaluate at each epoch, for 30 epochs.

B.1. f_s and f_t syntactic category MLPs

f_s and f_t are both MLPs with a linear input layer, two layers with ReLU non-linear activations, and finally an output linear layer. Their only difference is that the final linear layer output is either over non-terminal symbols or the vocabulary respectively.

B.2. Variational posterior model

In practice the variational posterior is given by a diagonal Gaussian where the mean and log-variance vectors are given by another biLSTM with a maxpooled linear output layer over the hidden states, for **z** over each batch of sentences.

Appendix C. Syntactic category mappings

POS tags from SpaCy were mapped to the following syntactic categories using this correspondences in Table 3.

Figs. 14 through 17 are the predicted category to syntactic category mappings for all other model runs.

Finally, Fig. 18 plots the mean Jensen–Shannon divergence between predicted syntactic categories across random seed runs for each model. The closer values are to zero the more similar category distributions are. For example, we note that the distributions for proper nouns and nouns are very close under the joint-learning models, while the same can be said for nouns and adjectives under the syntax-first model.

Data availability

I have shared a link to my code and data and it is all publicly available.

Table 3
Correspondence between syntactic categories and part-of-speech tags.

Syntactic category	Part-of-speech tags
Verb	VB, VBD, VBG, VBN, VBP, VBZ
Adjective	JJ
Adverb	RB, RBR, RP, RBS
Noun	NN, NNS
Proper noun	NNP, NNPS
Pronoun	PRP, PRP$
Determiner	DT
Conjunction	CC
Modal	MD
Other function word	IN

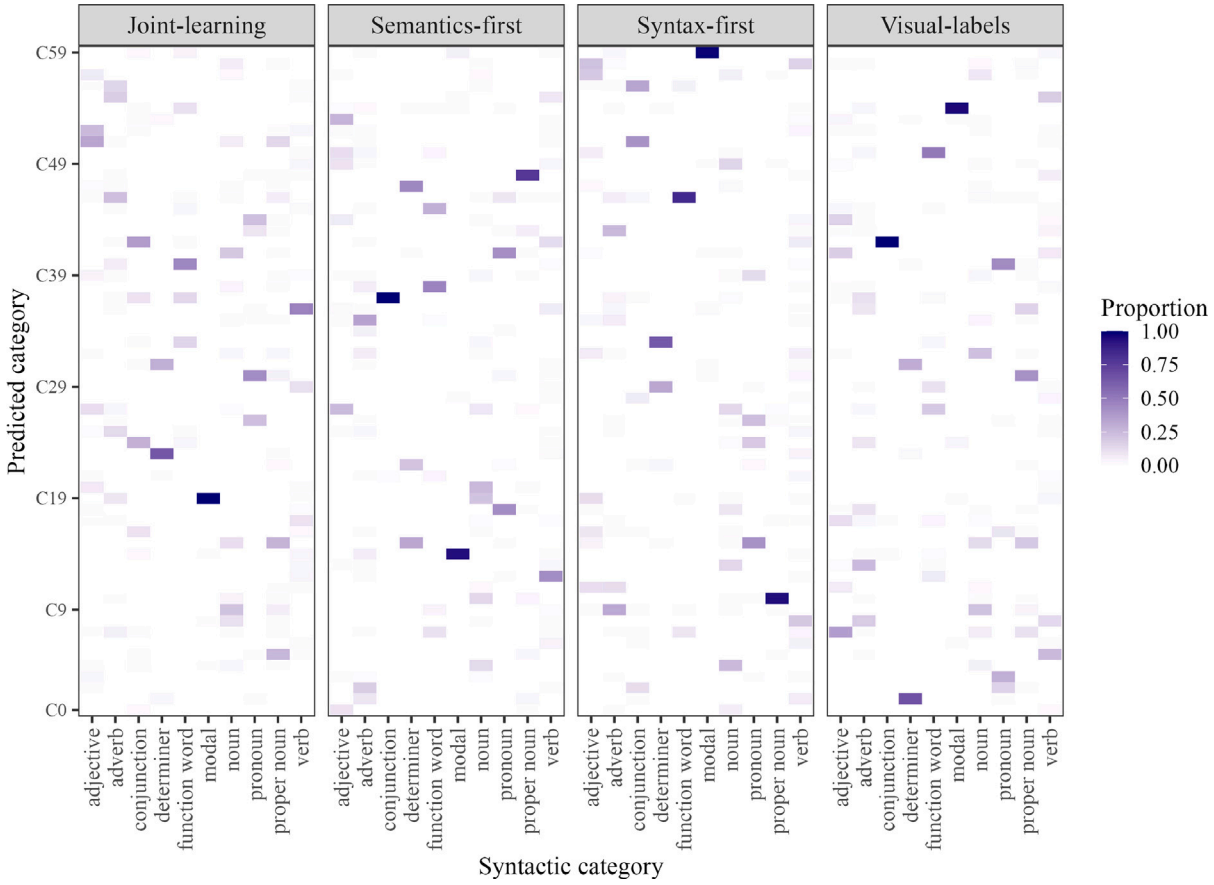


Fig. 14. Proportion of predicted category to syntactic category mappings on all sentences for models with random seed 214.

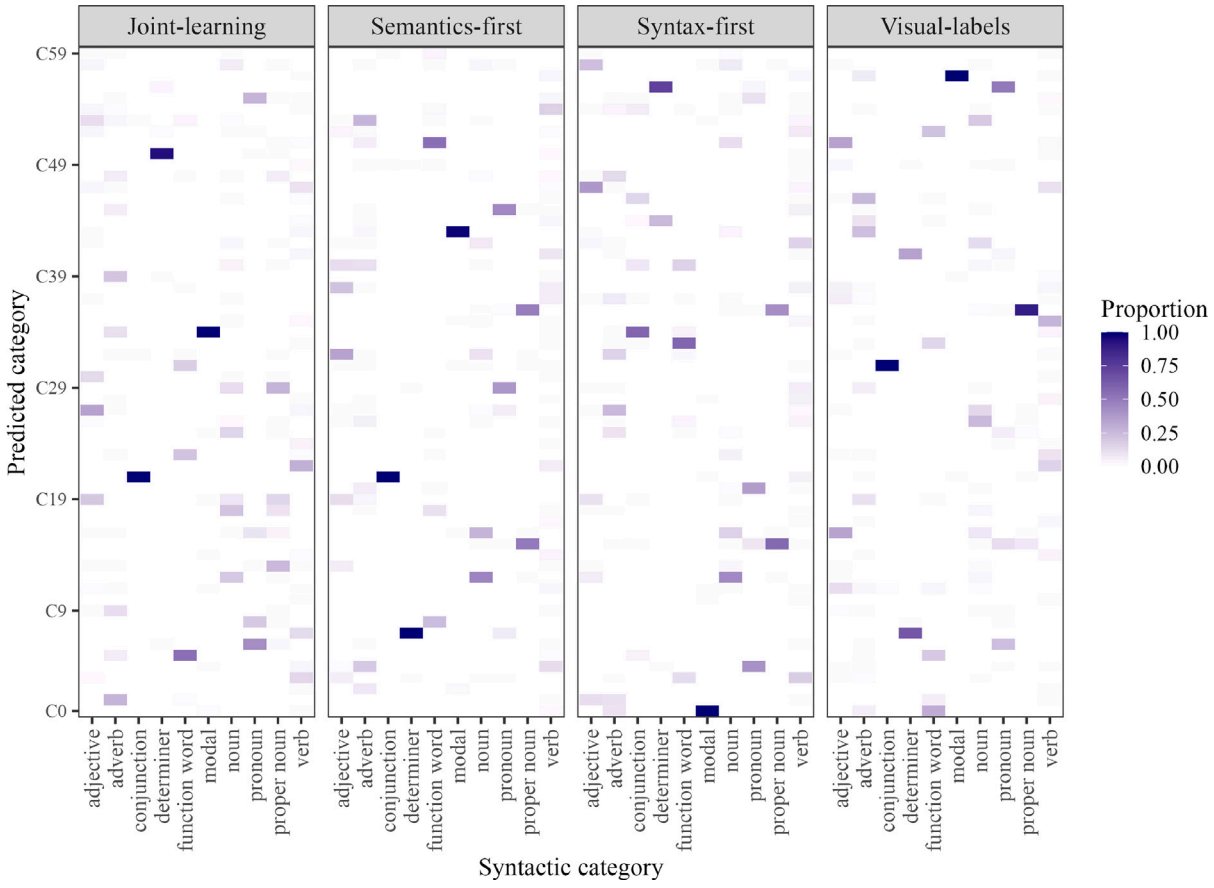


Fig. 15. Proportion of predicted category to syntactic category mappings on all sentences for models with random seed 527.

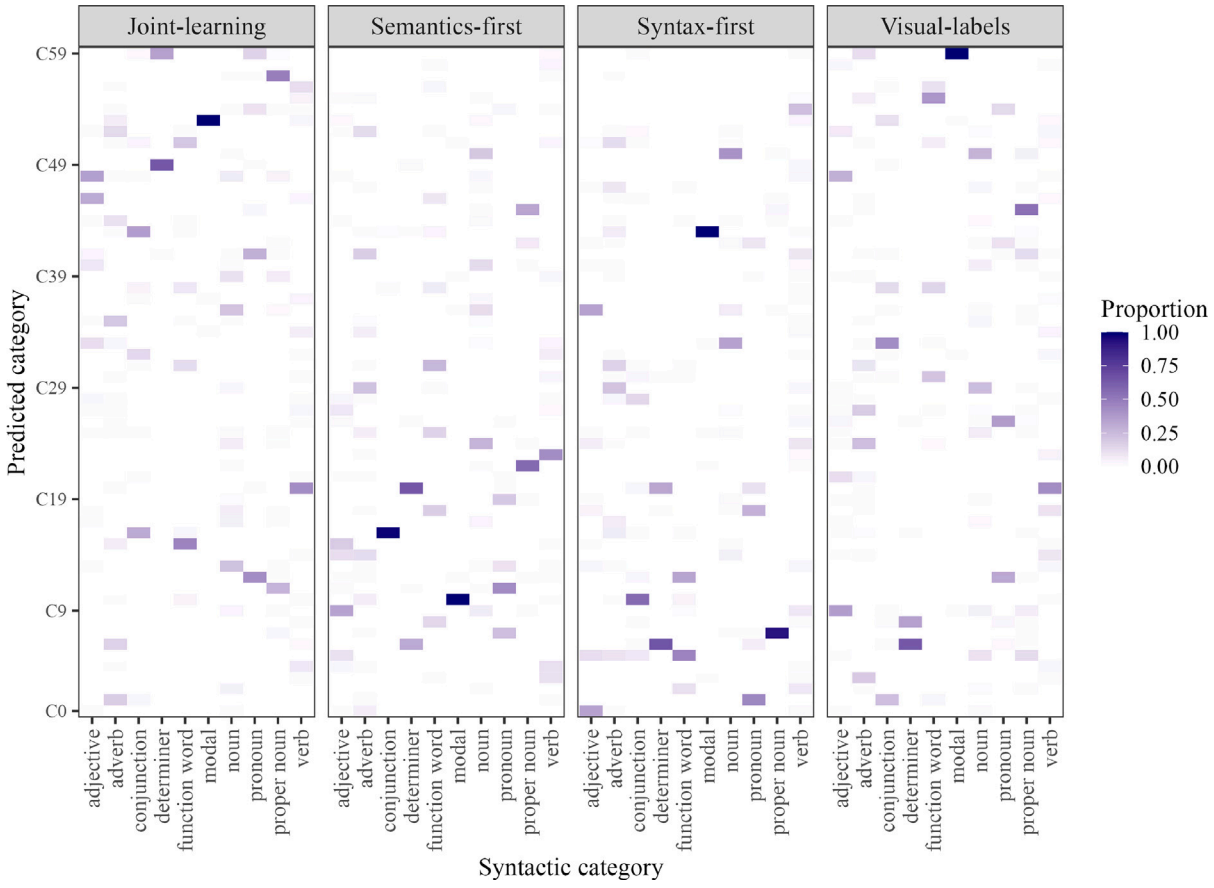


Fig. 16. Proportion of predicted category to syntactic category mappings on all sentences for models with random seed 627.

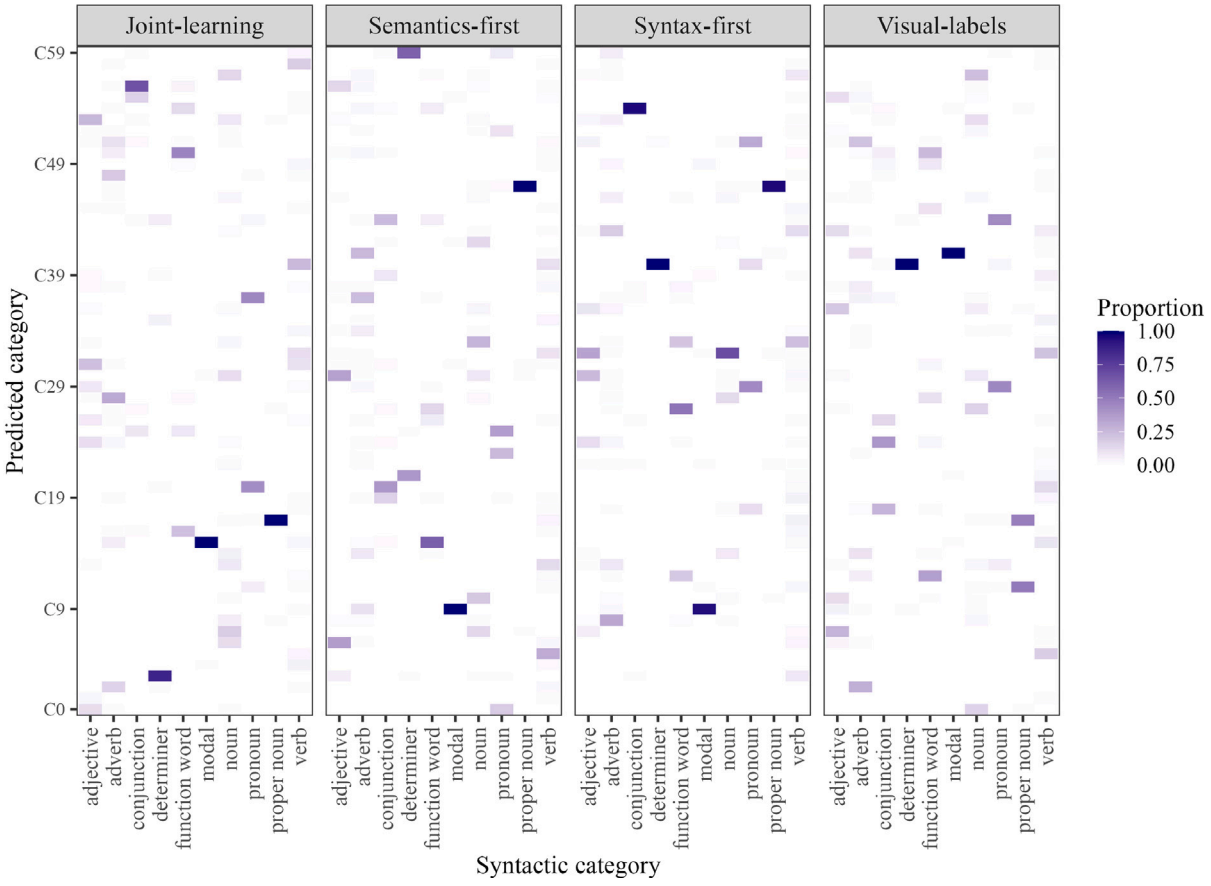


Fig. 17. Proportion of predicted category to syntactic category mappings on all sentences for models with random seed 91.

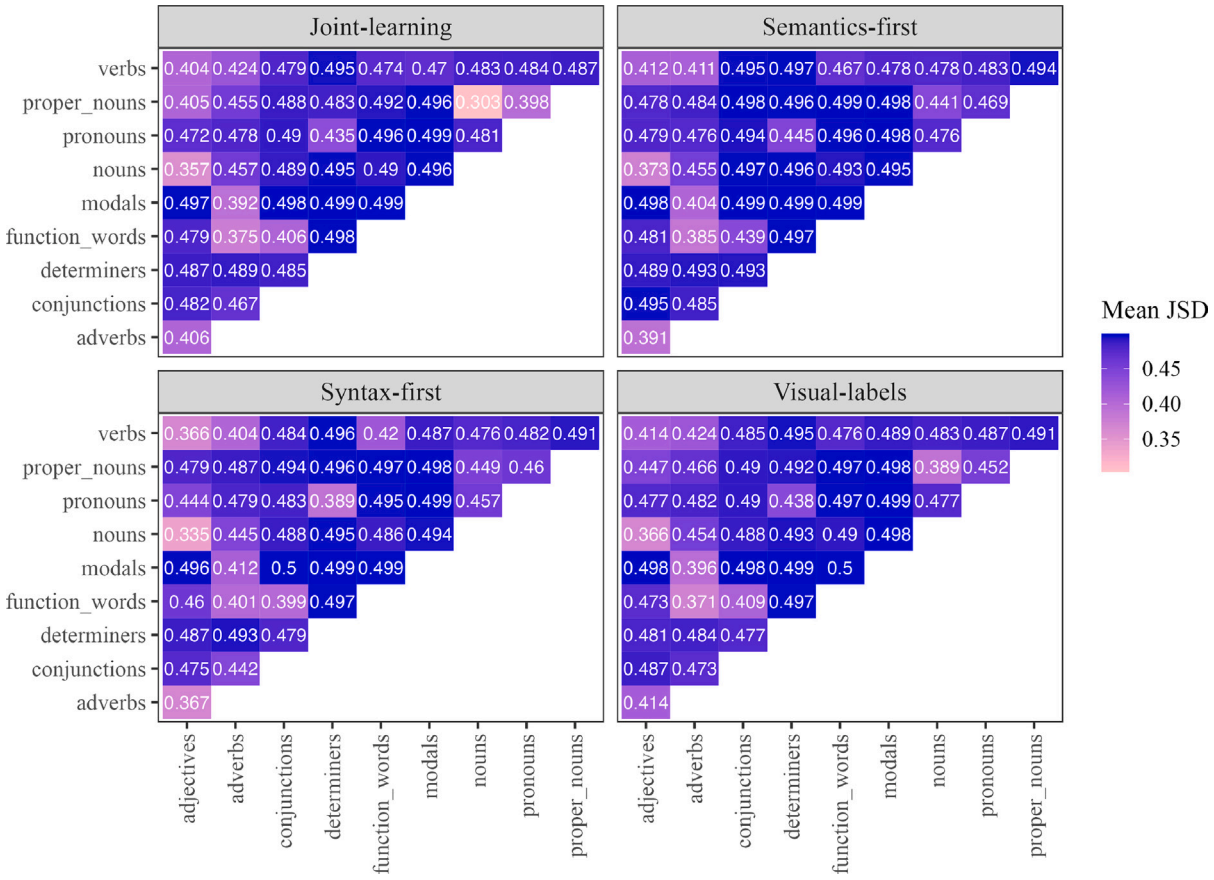


Fig. 18. Mean Jensen–Shannon divergence between predicted syntactic categories. The closer values are to zero the more similar category distributions are.

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., et al., Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Brinchmann, E. I., Braeken, J., & Lyster, S. A. H. (2019). Is there a direct relation between the development of vocabulary and grammar? *Developmental Science*, *22*(1), Article e12709.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344. http://dx.doi.org/10.1016/j.tics.2006.05.006.

Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., et al. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Clark, A., & Lappin, S. (2011). *Linguistic nativism and the poverty of the stimulus*. John Wiley & Sons.

Cohn, T., Blunsom, P., & Goldwater, S. (2010). Inducing tree-substitution grammars. *The Journal of Machine Learning Research*, *11*, 3053–3096.

Crain, S., & Thornton, R. (2012). Syntax acquisition. *WIREs Cognitive Science*, *3*(2), 185–203. http://dx.doi.org/10.1002/wcs.1158, https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1158.

Drozdov, A., Verga, P., Yadav, M., Iyyer, M., & McCallum, A. (2019). Unsupervised latent tree induction with deep inside-outside recursive auto-encoders. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)merican chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 1129–1141). Association for Computational Linguistics, https://aclanthology.org/N19-1116.

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, Article 104348.

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brainscape of the human brain. *Nature Reviews Neuroscience*, *25*, 289–312.

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 143–149.

Fisher, C., Jin, K. S., & Scott, R. M. (2020). The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science*, *12*(1), 48–77.

Frank, M. C., Braginsky, M., Marchman, V., & Yurovsky, D. (2021). *Variability and consistency in early language learning: the wordbank projectonsistency in early language learning: the wordbank project*. MIT Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23–64.

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C. L. Baker, & J. J. McCarthy (Eds.), *The logical problem of language acquisition* (pp. 183–210). MIT Press.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1725–1744). http://dx.doi.org/10.18653/v1/2020.acl-main.158, https://aclanthology.org/2020.acl-main.158/.

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., et al., Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., ... Fedorenko, E. (2022). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, *33*(8), 4384–4404. http://dx.doi.org/10.1093/cercor/bhac350.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning* coNLL, (pp. 624–646).

Huybregts, R. (1984). The weak inadequacy of context-free phrase structure grammars. In G. J. de Haan, M. Trommelen, & W. Zonneveld (Eds.), *Van periferie naar kern* (pp. 81–99). Foris Publications, Dordrecht.

Jin, L., & Schuler, W. (2020). Grounded PCFG induction with images induction with images. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing* (pp. 396–408). https://aclanthology.org/2020.aacl-main.42.

Kim, Y., Dyer, C., & Rush, A. (2019). Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2369–2385). Association for Computational Linguistics, https://aclanthology.org/P19-1228.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.

Kitaev, N., Cao, S., & Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3499–3505). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1340, https://www.aclweb.org/anthology/P19-1340.

Kitaev, N., & Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2676–2686). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1249, https://www.aclweb.org/anthology/P18-1249.

Klein, D., & Manning, D. C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd annual meeting of the association for computational linguistics* ACL-04, (pp. 478–485).

Klein, D., & Manning, D. C. (2005). Natural language grammar induction with a generative constituent-context model. *Pattern Recognition*, *38*(9), 1407–1419.

Köylü, Y. (2021). An overview of the NP versus DP debate. *Language and Linguistics Compass*, *15*(3), Article e12406.

Landau, B., & Gleitman, L. R. (1985). *Language and experience: evidence from the blind child*. Harvard University Press.

Lappin, S. (2021). *Deep learning and linguistic representation*. CRC Press.

Li, B., Corona, R., Mangalam, K., Chen, C., Flaherty, D., Belongie, S., et al., Li, B., Corona, R., Mangalam, K., Chen, C., Flaherty, D., Belongie, S., ... Klein, D. (2024). Re-evaluating the need for multimodal signals in unsupervised grammar induction.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al., Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the 13th European conference of computer vision* (pp. 740–755).

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependenciess to learn syntax-sensitive dependencies. *TACL*, *4*, 521–535.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, http://dx.doi.org/10.1016/j.tics.2024.01.011.

Muralidaran, V., Spasić, I., & Knight, D. (2020). A systematic review of unsupervised approaches to grammar induction. *Natural Language Engineering*, *27*(6), 647–689.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*(2), 357–374.

Nikolaus, M., & Fourtassi, A. (2021a). Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 200–210). https://aclanthology.org/2021.cmcl-1.24.

Nikolaus, M., & Fourtassi, A. (2021b). Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledges early acquisition of semantic knowledge. In *Proceedings of the 25th conference on computational natural language learning* (pp. 391–407). https://aclanthology.org/2021.conll-1.31.

Pearl, L. (2023). Computational cognitive modeling for syntactic acquisition: Approaches that integrate information from multiple places. *Journal of Child Language*, 1–21.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Piantadosi, S. (2023). Modern language models refute chomsky's approach to language. (p. 7180). Lingbuzz Preprint, lingbuzz.

Pinker, S. (1984). *Vol. 1, Language learnability and language development*. Harvard University Press.

Pinker, S. (2009). *Vol. 7, Language learnability and language development: with new commentary by the author*. Harvard University Press.

Portelance, E., & Jasbi, M. (2024). The roles of neural networks in language acquisition. *Language and Linguistics Compass*, *18*, Article e70001.

Portelance, E., & Jasbi, M. (2025). On the compatibility of generative AI and generative linguistics and generative linguistics. arXiv preprint arXiv:2411.10533.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al., Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the international conference on machine learning* (pp. 8748–8763).

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second berkeley symposium on mathematical statistics and probability* (pp. 131–149).

Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure-measure: A conditional entropy-based external cluster evaluation measure. In J. Eisner (Ed.), *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning* EMNLP-coNLL, (pp. 410–420). https://aclanthology.org/D07-1043.

Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., et al., Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., ... Fedorenko, E. (2024). Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience*, 1–43. http://dx.doi.org/10.1162/jocn_a_02164.

Shi, H., Mao, J., Gimpel, K., & Livescu, K. (2019). Visually grounded neural syntax acquisition. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1842–1861). https://aclanthology.org/P19-1180.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. In J. Kulas, J. H. Fetzer, & T. L. Rankin (Eds.), *Philosophy, language, and artificial intelligence: resources for processing natural language* (pp. 79–89). Springer Netherlands.

(2020). spaCy: Industrial-strength Natural Language Processing in Python.y: Industrial-strength. *Natural Language Processing in Python*, http://dx.doi.org/10.5281/zenodo.1212303.

Tsuji, S., Cristia, A., & Dupoux, E. (2021). SCALa: A blueprint for computational models of language acquisition in social context. *Cognition*, *213*, Article 104779.

Wan, B., Han, W., Zheng, Z., & Tuytelaars, T. (2022). Unsupervised vision-language grammar induction with shared structure modeling. In *International conference on learning representations*.

Warstadt, A., & Bowman, R. S. (2023). What artificial neural networks can tell us about human language acquisition. In S. Lappin, & J. P. Bernardy (Eds.), *Algebraic structures in natural language*. Taylor & Francis Group.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., et al. (2020). BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. http://dx.doi.org/10.1162/tacl_a_00321.

Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development*, *83*(4), 1382–1399.

Zhao, Y., & Titov, I. (2020). Visually grounded compound PCFGss. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 4369–4379). https://www.aclweb.org/anthology/2020.emnlp-main.354.

Zitnick, C. L., & Parikh, D. (2013). Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3009–3016). https://ieeexplore.ieee.org/document/6619231.

Zitnick, C. L., Parikh, D., & Vanderwende, L. (2013). Learning the visual interpretation of sentences. In *Proceedings of the IEEE international conference on computer vision* (pp. 1681–1688). https://ieeexplore.ieee.org/document/6751319.