Main clause syntax and the labeling problem in syntactic bootstrapping[*]

Aaron Steven White
Johns Hopkins University

Valentine Hacquard
University of Maryland

Jeffrey Lidz
University of Maryland

Abstract

In English, the distinction between belief verbs, such as *think*, and desire verbs, such as *want*, is tracked by the tense of those verbs' subordinate clauses. This has led some authors within the syntactic bootstrapping literature to propose that subordinate clause tense is an integral part of the acquisition of belief and desire predicates. This proposal is problematic since the correlation between tense and the belief v. desire distinction is not cross-linguistically robust, yet the acquisition profile for these verbs appears to be identical cross-linguistically. Thus, a story on which a particular syntactic feature—such as tense—cues the learner to the appropriate label for a particular semantic distinction—belief v. desire—will not work unmodified. Our proposal in this chapter is that, rather than being cued to a semantic distinction, like belief v. desire, by a particular syntactic feature, like subordinate clause tense, learners may utilize more abstract syntactic cues, whose instantiation is constrained to a small set of possible syntactic feature configurations, but must ultimately be tuned to the syntactic distinctions present in a particular language.

1 Introduction

Syntactic bootstrapping encompasses a family of approaches to verb learning wherein learners use the syntactic contexts a verb is found in to infer its meaning (Landau and Gleitman, 1985; Gleitman, 1990). Any such approach must solve two problems. First, it must specify how learners *cluster* verbs—i.e., figure out that some set of verbs shares some meaning component—according to their syntactic distributions. For instance, an English learner might cluster verbs based on whether they embed tensed subordinate clauses (1) or whether they take a noun phrase (2).

(1) a. John {thinks, believes, knows} that Mary is happy.
b. *John {wants, needs, orders} that Mary is happy.

(2) a. John {believes, knows, wants, needs} Mary.
b. *John {thinks, orders} Mary.

For different parts of the lexicon, different clusterings may be better or worse. Among propositional attitude verbs—like *think*, *believe*, *know*, *want*, *need*, and *order*—clustering based on whether the verb takes a subordinate clause yields intuitively better clusters than clustering based on whether it takes a noun phrase (at least when these structures are considered in isolation). That is, CLUSTERs 1 and 2 are intuitively more coherent than CLUSTERs 3 and 4.

(3) a. CLUSTER 1: think, believe, know

[*] We are grateful to audiences at Semantics and Linguistic Theory 24, the University of Maryland, and UC Santa Cruz as well as Alexander Williams, Shevaun Lewis, Kaitlyn Harrigan, Naho Orita, Rachel Dudley, Morgan Moyer, and Tom Roberts for useful comments. This work was supported by NSF BCS DDRIG grant 1456013, NSF BCS grant 1124338, and NSF DGE IGERT grant 0801465. This manuscript was submitted to *Semantics in Acquisition*, Trends in Language Acquisition Research (eds. K. Syrett and S. Arunachalam). Comments are welcome and should be addressed to Aaron Steven White at aswhite@jhu.edu.

b. CLUSTER 2: want, need, order
c. CLUSTER 3: believe, know, want, need
d. CLUSTER 4: think, order

We refer to the problem of choosing how to cluster verbs based on syntactic context as the *clustering problem*, and we call the learning mechanism that solves this problem—i.e., outputs clusters like those in (3)—the *clustering mechanism*.

The second problem a syntactic bootstrapping approach must solve regards how learners *label* the clusters output by a clustering mechanism—i.e., figure out what meaning component a particular cluster of verbs corresponds to. For instance, a common way of labeling CLUSTERs 1 and 2 is to say that all verbs in CLUSTER 1 have a BELIEF component and all verbs in CLUSTER 2 have a DESIRE component.

(4) a. CLUSTER 1 ⟷ BELIEF
b. CLUSTER 2 ⟷ DESIRE

We refer to this second problem, which is in many ways more difficult than the clustering problem, as the *labeling problem*, and we call the learning mechanism that solves this problem—i.e., labels the clusters output by the clustering mechanism—the *labeling mechanism*.

In this paper, we present evidence from the domain of propositional attitude verbs that previous labeling mechanisms are unsatisfactory both empirically and explanatorily, and we propose a novel labeling mechanism. We focus in particular on the distinction among propositional attitude verbs between belief verbs, like *think*, and desire verbs, like *want*.

In the next section, we discuss the two main approaches to solving the labeling problem that have been instantiated in the literature and show that both of these approaches are inadequate—either because they make incorrect predictions or because they make essentially no predictions at all. Throughout this section, we make reference to the belief-desire distinction as an exemplar, though we defer a detailed discussion of its distributional properties until Section 3. In Section 4, we present our proposal, which is a hybrid to the two previous approaches, at a conceptual level, then we present a computational level (Marr, 1982) description of the representations we employ in implementing the proposal. In Section 5, we implement this computational level description as a learning algorithm. In Section 6, we present a brief proof-of-concept experiment, which shows that our model finds a correct labeling when run on syntactic distributions found in child-direct speech (CDS). In Section 7, we conclude with some remarks on what our proposal entails for the theory of verb learning.

2 Approaches to the labeling problem

Current approaches to the labeling problem fall into two broad categories: the *top-down approach* and the *bottom-up approach*. In the top-down approach—the traditional one laid out in Landau and Gleitman 1985; Gleitman 1990—labeling is part-and-parcel with clustering. The learner has some innate mappings from semantic features to syntactic features (*projection rules*; Gruber 1965; Carter 1976; Chomsky 1981; Pinker 1984, 1989; Grimshaw 1990; Levin 1993; Hale and Keyser 2002), and upon noticing that a particular verb occurs with a particular syntactic feature, the learner reverses those projection rules to get from that syntactic context to that word's corresponding semantic components (Kako, 1997; Lidz et al., 2004; White, 2015).

Continuing the examples above, a verb's taking a tensed subordinate clause correlates (in English) with that verb being a belief verb (5a), like *think*, *believe*, or *know* (Bolinger, 1968; Stalnaker, 1984; Farkas, 1985; Heim, 1992; Villalta, 2000, 2008; Anand and Hacquard, 2013, among others). This is corroborated by the fact that desire verbs, like *want*, *prefer*, and *order*, which arguably do not have a belief component, do not take tensed subordinate clauses (5b).[1]

(5) a. John {thinks, believes, knows} that Mary is happy.
b. *John {wants, needs, orders} that Mary is happy.

Assuming this correlation to be cross-linguistically robust, a syntactic bootstrapping account might then posit that learners have some innate projection rule (6a) that they can reverse to get from the fact that *think*, *believe*, and *know* occur with finite complements (6b) to the fact that *think*, *believe*, and *know* have a BELIEF meaning (6c) (De Villiers and De Villiers, 2000; De Villiers and Pyers, 2002; de Villiers, 2005).

(6) **Top-down approach**
a. *Knowledge:* BELIEF ⟶ S[+TENSE]
b. *Data:* {think, believe, know} S[+TENSE]
c. *Inference:* BELIEF ⟵ {think, believe, know}

The top-down approach makes strong predictions about learners' inferences: labeling is an automatic consequence of noticing a distributional fact—in this case, that a verb takes a tensed subordinate clause.

One difficulty that arises with the top-down approach is that it is not robust to cross-linguistic variation. For instance, suppose that the projection rule in (6a) were innate. One would expect either (i) that all languages show a correlation between a verb's having a BELIEF component and its taking tensed clauses; or (ii) that, if a language does allow non-belief verbs—e.g., desire verbs like *want*—to take tensed clauses, learners might go through a stage where they incorrectly believe that those verbs actually have a BELIEF component. Neither of these possibilities are realized: (i) there are languages, such as German, where both belief and desire verbs take tensed subordinate clauses (7); and (ii) in these languages, children do not mistake one type of verb for the other (Perner et al., 2003). (We discuss both of these points further in Section 3.)

(7) a. Ich glaube, dass Peter nach Hause geht.
I think that Peter to home goes.
b. Ich will, dass Peter nach Hause geht.
I want that Peter to home goes.

The bottom-up approach remedies this issue at the cost of making weaker distributional and developmental predictions. In the bottom-up approach (Alishahi and Stevenson, 2008), learners cluster verbs (8c-i) based on syntactic context (8b), but the clustering mechanism itself does not provide the labels for these clusters. Rather, the learner must notice some correlation between the unlabeled clusters and the sorts of conceptualizations they have when that cluster is instantiated (8c-ii).[2] Then, given the cluster each verb falls into (8c-i) and the labeling of that cluster (8c-ii), the learner can make the inference that those verbs have that label (8c-iv).

(8) **Bottom-up approach**
a. *Knowledge:* ∅[3]

[1] But see Heim (1992) for a proposal on which *want* does have a belief component.

[2] The bottom-up approach is similar in form to semantic bootstrapping, in which learners are presumed to have access to the semantics relevant to a particular learning instance (Grimshaw, 1981, 1994; Pinker, 1984, 1989, 1994): in this case, the fact that the conceptual content BELIEF "cooccurs" with the linguistic content {think, believe, know} S[+TENSE]. It differs, however, in the sense that semantic bootstrapping is a theory of how children come to learn the syntax of their language, whereas the bottom-up approach assumes access to the syntax as a prerequisite. Further, the traditional version of the semantic bootstrapping is more like the top-down approach in assuming learners have innate projection rules (though see Connor et al. 2013).

[3] There is of course knowledge that learners are required to have for either approach to work that we are not

b. *Data:* (BELIEF, {think, believe, know} S[+TENSE])

c. *Inferences*

   (i) CLUSTER 1 $\longleftarrow$ {think, believe, know}

   (ii) CLUSTER 1 $\longleftrightarrow$ BELIEF

   (iii) CLUSTER 1 $\longleftrightarrow$ S[+TENSE]

   (iv) BELIEF $\longleftarrow$ {think, believe, know}

   (v) BELIEF $\longrightarrow$ S[+TENSE]

Importantly, the bottom-up approach can also learn the projection rule from BELIEF to tense (8c-v) by noticing a correlation between the cluster and the syntax (8c-iii)—similar to the explanation for how the label itself is associated with the cluster. Indeed, some bottom-up models, such as Alishahi and Stevenson's (2008), explicitly treat the association between the cluster and the concept as of the same type as the association between the cluster and the syntactic feature—i.e., the projection rule. This is because they treat both the concept and the syntax as observed features of the verb, which can both be used in forming the cluster in the first place.

The bottom-up approach has two major problems. First, it makes weak predictions about learners' inferences. Since the labeling mechanism is a separate component of the overall word-learning mechanism, learners could in principle learn any relationship between unlabeled cluster, conceptual label, and syntactic feature—not just cross-linguistically attested ones. Relatedly, it makes no predictions about what should be cross-linguistically attested. Since the bottom-up approach has no way of encoding biases for particular concept-to-syntax mappings, it has no way of explaining correlations between the syntax and the semantics, even if it has a way of learning them.

Second, the bottom-up approach requires the rejection of one of the main arguments for syntactic bootstrapping in the first place: the fact that some semantic components seem to be "closed to observation"—the parade case being components associated with abstract conceptual content like BELIEF and DESIRE (Gleitman, 1990; Gillette et al., 1999; Snedeker and Gleitman, 2004; Gleitman et al., 2005; Papafragou et al., 2007). One cannot observe believing or wanting.

Both of these weaknesses remain unremedied in recent instantiations of the bottom-up approach that focus explicitly on propositional attitude verbs (Barak et al., 2012, 2013, 2014a,b). In fact, for the purposes of learning a word's meaning, the syntactic features are to some extent superfluous for models like Alishahi and Stevenson's and Barak et al.'s, since the semantics themselves are observed and can thus contribute to forming a cluster with a particular label. In this sense, the bottom-up approach is essentially a cross-situational word learning model (Yu and Smith, 2007; Smith and Yu, 2008; Yu and Smith, 2012; Medina et al., 2011; Trueswell et al., 2013) with additional features (cf. Frank et al., 2009). But for the same reason that the bottom-up approach is explanatorily weak, it is robust to cross-linguistic variability. As long as some syntactic feature correlates with a particular semantic distinction, the relationship between that feature and that distinction should be learnable.

This robustness to cross-linguistic variability comes with a prediction: any correlation between a syntactic feature and a semantic distinction is purely a coincidence of a particular language's lexicon, at least from the learner's point-of-view. Thus, evidence for an interpretable set of constraints on how a semantic feature relates to a syntactic feature is evidence against a bottom-up approach. Such evidence does not obviate the fact that the top-down approach cannot work unmodified, but as we show below, it can usefully guide us toward an approach with similar explanatory power.

listing here. For example, both the top-down and bottom-up approaches require (i) that the relevant syntactic structures can be parsed by learners at the relevant developmental stage, and (ii) that the relevant conceptual material is accessible to them at that stage. This second requirement may or may not be met at certain points in development. See Onishi and Baillargeon 2005; Baillargeon et al. 2010 for evidence that this conceptual material is accessible from a very young age.

In the next section, we delve into the cross-linguistic distributional facts pertaining to the belief-desire distinction. We give evidence (i) that the belief-desire distinction is not tracked by a stable set of syntactic features across languages but (ii) that this instability is highly constrained. In particular, we suggest that the syntactic features that correlate with the distinction in a particular language are exactly those features found in that language's main clauses. We argue that, stated in this way, the correlation between the belief-desire distinction and the syntax of a particular language is interpretable as a consequence of a semanticopragmatic property of particular belief predicates known as *assertivity* (Hooper, 1975).

3 The belief-desire distinction across languages

In this section, we briefly discuss attitude verb classifications relevant to the belief-desire distinction along with the cross-linguistic syntactic correlates of that distinction. As noted above, the belief-desire distinction—or rather, a related, but more general, distinction between representational verbs and preferential verbs—is cross-linguistically attested, but the syntactic correlates of this distinction are unstable. In spite of this instability, we see that these syntactic correlates all have a family resemblance: these syntactic correlates are just the ones that distinguish declarative matrix clauses from other clause types. We then discuss the semanticopragmatic reasons why such a correlation should exist. We begin by reviewing the known syntactic correlates of the belief-desire distinction and then discuss the semanticopragmatic property—*assertivity*—that likely drives this correlation.

3.1 Representationals and preferentials in English

Propositional attitude verbs have highly multifaceted semantic representations—i.e., these verbs can be clustered into many cross-cutting groups (see White 2015 for a review). One of the most well-studied facets of attitude verb meaning is the distinction between verbs that express beliefs—or more generally, represent "mental pictures" or "judgments of truth" (Bolinger, 1968)—and those that express desires—or more generally, preferences for particular states of affairs associated with, e.g. commands, laws, preferences, etc. (Bolinger, 1968; Stalnaker, 1984; Farkas, 1985; Heim, 1992; Villalta, 2000, 2008; Anand and Hacquard, 2013, a.o.). Within the first class, which Bolinger (1968) calls the *representationals*, fall belief verbs like *think*, as well as communication verbs, like *say*; and within the second class, which we refer to as the *preferentials*, fall verbs like *want* and *order*.

In English, the distinction between representationals and preferentials is correlated strongly with subordinate clause tense. Representationals tend to allow tensed subordinate clauses (9a) while preferentials tend not to (9b).

(9) a. John thinks that Mary went to the store.
b. *John wants that Mary went to the store.

This generalization is often strengthened to incorporate the converse: representationals tend not to allow untensed subordinate clauses (10a) while preferentials tend to (10b).

(10) a. *John thinks Mary to go to the store.
b. John wants Mary to go to the store.

The stronger version of this statement is supported by the existence of verbs that appear to have both representational and preferential semantic components and can take both tensed and untensed complements. For instance, *hope p* involves both a desire that *p* come about and the belief that *p* is possible (Portner, 1992; Scheffler, 2009; Anand and Hacquard, 2013; Hacquard, 2014; Harrigan, 2015, but see also Portner and Rubinstein 2013), and it occurs in both finite (11a) and nonfinite (11b) syntactic contexts. (This is true more generally for *emotive* verbs like *wish*, *fear*, etc., though the exact form of the untensed complement may differ.)

(11) a. John hopes that Mary went to the store.
b. John hopes to go to the store.

The correlation between subordinate clause tense and the distinction between representationals and preferentials is imperfect (even in English). For example, both *believe* and *remember* can take untensed subordinate clauses. This suggests that the stronger version of the tense generalization—that representationals do not take untensed subordinate clauses—is imperfect.

(12) a. John believes Mary to be intelligent.
b. John remembered to take out the trash.

Further, some preferential verbs appear to allow subordinate clauses that look similar to the tensed subordinate clauses that representationals take. For instance, *demand* can take clausal complements headed by *that*, suggesting that even the weaker form of the generalization is imperfect.

(13) John demanded that Mary go to the store.

This datum is perhaps less damaging to the weaker form of the generalization, since the kind of *that* clauses that *demand* and other preferentials take generally require the verb to be in the bare form. This could mean that these clauses are in fact untensed or that they involve a distinct mood (for speakers of a dialect with the English subjunctive). In any case, these sorts of subordinate clauses are extremely rare in child-directed speech—perhaps even nonexistent for many learners—so at least the weaker form of the generalization seems safe for the purposes of learning.

Beyond subordinate clause tense, the distinction between representationals and preferentials appears to correlate with whether the verb's subordinate clause can be fronted—or in Ross's (1973) terms, S-lifted. At least some representationals' subordinate clauses (14) appear to be able to undergo S-lifting, but many preferentials' subordinate clauses (15) cannot (Bolinger, 1968). (Not all representationals allow S-lifting—likely because the availability of S-lifting for a particular verb is conditioned by other semantic and pragmatic properties it has.)

(14) Mary already went to the store, I {think, believe, suppose, hear, see}

(15) a. *John already went to the store, I {want, need, demand}.
b. *John to go to the store, I {want, need, order}.

The ability to occur in S-lifting constructions appears to be related to the ability to allow complementizer drop and the propositional anaphor *so*: many verbs that allow S-lifting also allow complementizer drop and anaphoric *so* (Hooper, 1975; Grimshaw, 2009).

(16) a. I {think, know} (that) Jo already went to the store.
b. I {hate, love} *(that) Jo already went to the store.

The availability of these structures is related to a semanticopragmatic verb class Hooper (1975) calls the *assertives*. A verb is assertive if it can be used by speakers to make indirect assertions, as can be seen in (17a) (see Urmson 1952; Simons 2007; Lewis 2013; Anand and Hacquard 2014 for discussion). For instance, *think* and *say* seem to allow this (17a), but *hate* does not (17b).

(17) a. **A:** Where is Mary?
B: John {thinks, said} that she's in Florida.

b. **A:** Where is Mary?
B: # John hates that she's in Florida.

Assertivity is intimately tied with representationality in that the verbs that can show up in contexts like (17a) are always representational.[4] The availability of this use is likely due to the fact that these verbs express judgments of truth. This can be seen in the contrast between (18a) and (18b) (Stalnaker, 1984).

(18) a. John thinks it's raining, which is true.
b. John wants it to rain, #which is true.

We see in (18a) that the speaker can indirectly assert the content of the complement clause, by endorsing the reported judgment of truth. In contrast, desire verbs express preferences, and they are not routinely used to make indirect assertions. Instead, they are often used to make indirect requests. For instance, sentences like (19) can be used to express more than a mere desire, but to make a request.

(19) I want you to go to your room!

Thus, we see that the assertive verbs, which are a subset of the representational verbs are also verbs that take tensed subordinate clauses whose complementizer can be dropped. This means that representationals' embedded clauses can match the syntactic features of a corresponding declarative main clause in English.

(20) It's raining.

We show in the next section that, even in the face of cross-linguistic variation in which syntactic features correlate with the distinction between representationals and preferentials, this match between representational subordinate clauses and declarative main clauses remains constant.

3.2 Representationals and preferentials in other languages

The syntax-semantics correlations discussed above do not necessarily hold cross-linguistically. In some languages, the distinction is roughly tracked by mood. In the Romance languages, representationals tend to take indicative mood and preferentials tend to take subjunctive mood (Bolinger, 1968; Hooper, 1975; Farkas, 1985; Portner, 1992; Giorgi and Pianesi, 1997; Giannakidou, 1997; Quer, 1998; Villalta, 2000, 2008; Anand and Hacquard, 2013, a.o.). For instance, in Spanish both the representational (belief) verb *creer* (*think/believe*) and the preferential (desire) verb *querer* (*want*) take finite subordinate clauses.

(21) a. Creo que Peter va a la casa.
think.1S.PRES that Peter go.PRES.IND to the house.
b. Quiero que Peter vaya a la casa.
want.1S.PRES that Peter go.PRES.SBJ to the house.

The difference between these subordinate clauses is that, whereas verbs like *creer* (*think*) take subordinate clauses with verbs inflected for indicative mood (21a), verbs like *querer* (*want*) take subordinate clauses with verbs inflected for subjunctive mood (21b). And as in English, the subordinate clause under the representational verb *creer* is featurally similar to the declarative main clause in Spanish, whose tensed verb is inflected for indicative mood.

(22) Peter va a la casa.
Peter go.PRES.IND to the house.

[4] Simons (2007) notes some potential counterexamples, but the contexts that these counterexamples can occur in are extremely circumscribed, suggesting they may not involve true indirect assertions.

Another example comes from some Germanic languages. As in Spanish, both representationals and preferentials in German allow tensed subordinate clauses.

(23) a. Ich glaube, dass Peter nach Hause geht.
I think that Peter to home goes.
b. Ich will, dass Peter nach Hause geht.
I want that Peter to home goes.

In subordinate clauses headed by the complementizer *dass* (*that*), the embedded verb occurs clause-finally, which evidences the fact that German is underlyingly a subject-object-verb (SOV) language. Both the verb *glauben* (*think*) and the verb *wollen* (*want*) can take such clauses, in which the main verb is tensed.

Only representationals like *glauben* (*think*), however, allow a second sort of structure known as verb second (V2) (Scheffler, 2009). If the complementizer *dass* (*that)* is not present, *glauben* (*think*) can take a subordinate clause with V2 syntax (24a). *Wollen* does not allow this (24b).

(24) a. Ich glaube, Peter geht nach Hause.
I think Peter goes to home.
b. *Ich will, Peter geht nach Hause.
I want Peter goes to home.

Thus, the distinction between representationals and preferentials is tracked by the availability of verb second (V2) syntax in their subordinate clauses (Truckenbrodt, 2006; Scheffler, 2009). V2 is furthermore found in German declarative main clauses. For instance, (25) shows a German main clause with the tensed form of the auxiliary verb *sein* (*be*) occurring as the second word of the sentence (in second position). Compare this to the subordinate clause in (24a).

(25) Peter geht nach Hause.
Peter goes to home

Thus, preferentials in both Spanish and German take tensed complements, militating against a hard-coded link between tense and representationality. But as we have been suggesting, they still show language-internal correlations between representationality and some more abstract aspect of the clausal syntax. The aspect of the clausal syntax that occurs with only the representational verbs—indicative mood in Spanish and V2 in German—also tends to show up in declarative main clauses (Dayal and Grimshaw, 2009; Hacquard, 2014).

3.3 Discussion

In this section, we noted a problem for the more explanatory top-down approach to syntactic bootstrapping: not all languages make use of subordinate clause tense to distinguish belief and desire verbs. However, they do seem to make the same broad distinctions (clusters), albeit via different syntactic features—Romance uses mood and German uses word order. We then suggested that, while the language specific syntactic features are different cross-linguistically, they all converge at a more abstract level: belief verbs take complements that have syntactic hallmarks of declarative main clauses—finite clauses in English, indicative mood in Romance, verb second in German.

We hypothesize that children can exploit the syntactic parallels between direct and indirect speech acts to reconstruct the underlying semantics. A verb like *think* takes complements with syntactic hallmarks of declarative main clauses, and is used to make indirect assertions. Thus, its meaning must be one that easily lends itself to indirect assertions: it must express a judgment of truth, which the speaker can endorse—i.e., it must have a representational semantics. We test the viability of such a proposal by asking whether a learner can discover the representational class by tracking whether the syntactic features of a verb's complements match those of the declarative main clause in their language.

4 Our proposal

In this section, we propose a model that exploits the fact that languages show a correlation between BELIEF semantic components and declarative main clause syntax. The two main contributions that we make over and above previous accounts are the concept of an *abstract projection rule* and the concept of a *featural anchor* for that projection rule. An abstract projection rule is a generalization of the traditional notion of a projection rule discussed in Section 2, wherein particular semantic components map onto unvalued syntactic features that must become valued over the course of learning. The featural anchor for that abstract projection rule is a class of syntactic structures that (i) determine how the syntactic features in the abstract projection rule are valued and (ii) are identifiable prior to verb learning. We suggest that the abstract projection rule for BELIEF has the class of declarative main clauses as its featural anchor.

4.1 The main conceptual components

In Section 2, we argued that the top-down (traditional) approach to syntactic bootstrapping is highly explanatory but brittle to cross-linguistic variability while the bottom-up approach is robust to cross-linguistic variability but has essentially no explanatory power. Ideally, there would be some way of combining the explanatory strengths of the top-down approach with the cross-linguistic robustness of the bottom-up approach. This is what our proposal aims to do. To carry this out, we propose to retain the notion of innate projection rules but incorporate a notion of abstract syntactic feature. So where the top-down approach has projection rules like (26a)—which we refer to as *concrete projection rules*—our model has projection rules like that (26b)—which we refer to as *abstract projection rules*.

(26) a. BELIEF $\longrightarrow$ S[+TENSE]
b. BELIEF $\longrightarrow$ MAIN CLAUSE
MAIN CLAUSE $\in$ S[+/-V2, +/-COMPLEMENTIZER, +/-MOOD, +/-TENSE]

This model lies mid-way between the top-down and bottom-up approaches in the following sense: like the top-down approach, it retains projection rules and thus the basis for a built-in labeling component, but like the bottom-up approach, it provides the flexibility to fit different languages, since different languages value the syntactic features in (26b) differently.

Indeed, our model can mimic the behavior of both approaches depending on how abstract the projection rules are. The projection rule (26a) can be thought of as a special case of an abstract projection rule in that concrete projection rules are just the limiting case of an abstract projection rule. In contrast, as the projection rules becomes more and more abstract—i.e., incorporate more and more features—they becomes essentially useless for labeling, thus requiring an alternative like the cross-situational learning found in the bottom-up approach.

Because our proposal can mimic the behavior of models that take either the top-down or bottom-up approach, there are two major hurdles our model faces. First, it must specify a set of syntactic features for each projection rule that is suitably broad to capture cross-linguistic facts but which is not so broad as to make the projection rule useless. Second, it must explain how the mechanism sets those features—i.e., turns the +/-s in (26b) into +s and −s.

We propose that the solution to these two problems is always based on the following principle: for every abstract projection rule, there corresponds a *featural anchor*. A featural anchor for a projection rule is some privileged class of linguistic structures that (i) instantiates all of the features in that rule and (ii) is identifiable without reference to those features.

4.2 A computational level description

What are the formal commitments of our proposal? We require at least five types of cognitive objects: (i) a set of verbs—e.g., {*think*, *want*, . . .}; (ii) a set of syntactic features—e.g., {TENSE, MOOD, . . .}; (iii) a set of semantic components—e.g., {BELIEF, DESIRE, . . .}; (iv) a set of projection rules (either abstract or concrete); and (v) a featural anchor (a set of valued syntactic features) for each abstract projection rule.[5] Using these objects, the learner's job is then to link each verb with a conceptual/semantic representation or some set thereof.

For the sake of exposition, we will also consider a sixth sort of object: an indexed set of variables over semantic components. This set models the unlabeled clusters one sees in the bottom-up approach, wherein the cluster index stands as a placeholder for a semantic component to be provided by the labeling mechanism. We incorporate it here since it allows us to show how our hybrid model differs from a bottom-up approach mathematically, though it is not a necessary component of our proposal. Indeed, we believe that minimizing or jettisoning the use of such variables is a criterion for explanatory adequacy—since ideally, one can retain the very strong hypothesis present in the top-down approach that the clustering and labeling components are one and the same thing (at least insofar as the use of syntax as a word-learning cue is concerned).

Let's begin by first setting up how we represent the relationship between verbs and semantic components. To a large extent, these representations are those proposed by White et al. (under reviewa) in his computational model of syntactic bootstrapping; however, our proposal can be implemented in, e.g., Alishahi and Stevenson's (2008) with few changes. We prefer to work with White et al.'s representations since they more transparently map to traditional notions in syntactic and semantic theory.

Given the existence of verbs like *hope*, which have both BELIEF and DESIRE components, we assume that each verb is linked with a (possibly singleton) set of semantic representations. For instance, if BELIEF and DESIRE were the only semantic representations, the English adult state would look like (27).

(27) a. think → {BELIEF}
b. hope → {BELIEF, DESIRE}
c. want → {DESIRE}

Note that for any fixed set of semantic components, the mappings in (27) can be represented as vectors of ones and zeros (Stone, 1936). So for instance, if BELIEF and DESIRE were again the only semantic representations, we could equivalently represent (27) as in (28), where the first coordinate corresponds to BELIEF and the second coordinate corresponds to DESIRE.

(28) a. think → $\begin{array}{cccc} & \text{BELIEF} & \text{DESIRE} & \\ (& 1 & 0 &) \end{array}$
b. hope → $\begin{array}{cccc} & \text{BELIEF} & \text{DESIRE} & \\ (& 1 & 1 &) \end{array}$
c. want → $\begin{array}{cccc} & \text{BELIEF} & \text{DESIRE} & \\ (& 0 & 1 &) \end{array}$

This suggests a compact way of representing semantic components in the lexicon; we bundle each vector in (28) into a matrix, where rows of the matrix correspond to verbs and columns correspond to semantic components. We'll call this matrix **S** for semantic components.

[5] There is an important distinction between conceptual representations and semantic representations, but this distinction is immaterial here.

(29) $$\mathbf{S} = \begin{array}{r} \\ \text{think} \\ \text{hope} \\ \text{want} \end{array} \begin{array}{cc} \textsc{belief} & \textsc{desire} \\ \left(\begin{array}{c} 1 \\ 1 \\ 0 \end{array}\right. & \left.\begin{array}{c} 0 \\ 1 \\ 1 \end{array}\right) \end{array}$$

This in turn provides a natural way to state the labeling problem. First, note that the setup so far just concerns the representation of which verbs have which semantic components. Every theory of syntactic bootstrapping, regardless of approach, needs some way of doing this. To state the labeling problem, simply replace BELIEF and DESIRE in (29) with CLUSTER 1 and CLUSTER 2. Let's call the resulting matrix $\mathbf{S}'$.

(30) $$\mathbf{S}' = \begin{array}{r} \\ \text{think} \\ \text{hope} \\ \text{want} \end{array} \begin{array}{cc} \textsc{cluster } 1 & \textsc{cluster } 2 \\ \left(\begin{array}{c} 1 \\ 1 \\ 0 \end{array}\right. & \left.\begin{array}{c} 0 \\ 1 \\ 1 \end{array}\right) \end{array}$$

The labeling problem is then stated as: how do learners map from $\mathbf{S}'$ to $\mathbf{S}$? Or focusing on just the column labels, how do they map from the labels on the columns of $\mathbf{S}'$, $\text{label}(\mathbf{S}') = (\textsc{cluster } 1, \textsc{cluster } 2)$ to the labels on the columns of $\mathbf{S}$, $\text{label}(\mathbf{S}) = (\textsc{belief}, \textsc{desire})$, possibly with the help of the ones and zeros in the matrix itself?

Let's call this mapping L (for labeling mechanism), where the domain and codomain have the same cardinality. (We'll also assume that this mapping is a function—i.e., maps each unlabeled cluster onto only a single semantic component—though nothing hinges on this.)

(31) $L : \{\textsc{cluster } 1, \textsc{cluster } 2, \ldots\} \rightarrow \{\textsc{belief}, \textsc{desire}, \ldots\}$

So how do the two approaches discussed above solve the labeling problem stated as such? That is, how do they select a mapping L? In the bottom-up approach, this is done using a separate source of data: pairings of unlabeled clusters with semantic components. Within the current setup, bottom-up approaches consider all possible functions from $\{\textsc{cluster } 1, \textsc{cluster } 2, \ldots\}$ to $\{\textsc{belief}, \textsc{desire}, \ldots\}$, since they have no idea *a priori* what the data will look like. This consideration of all possible functions is what gives rise to the dissatisfying aspect of this approach that no particular mapping is singled out.

In contrast, the top-down approach considers only one mapping L. Or perhaps, more parsimoniously, it has no notion of L, since it needs no notion of unlabeled cluster in the first place. To see this, it is useful to introduce a formalization of the projection rules. And to do this, it's useful to return to our concrete rule for BELIEF, repeated in (32a). This rule—and by definition, all projection rules—has the general form found in (32b).

(32) a. BELIEF ⟶ S[+TENSE]
b. SEMANTIC COMPONENT ⟶ SYNTACTIC CONFIGURATION
SYNTACTIC CONFIGURATION ∈ [+/-SYNTACTIC FEATURE 1, . . .]

We can then play (almost) the same encoding game we did for the semantic components. Each valued syntactic feature on the right side of a particular projection rule goes in a set corresponding to the semantic component.[6]

[6] One consequence of this encoding is that projection rules only select features they like; they never specify which features that dislike by, e.g., encoding those features by having a negative value. So, if DESIRE projected onto [-TENSE], the representation of the rule would be as in (ia)—not as in (ib) or (ic).

(i) a. $$\textsc{desire} \rightarrow \begin{array}{ccc} \textsc{+tense} & \textsc{-tense} & \cdots \\ \left(\ 0\right. & 1 & \left.\cdots\ \right) \end{array}$$

b. $$\textsc{desire} \rightarrow \begin{array}{ccc} \textsc{+tense} & \textsc{-tense} & \cdots \\ \left(\ -1\right. & 1 & \left.\cdots\ \right) \end{array}$$

(33) $$\begin{array}{rcccc} & & \text{+TENSE} & \text{-TENSE} & \cdots \\ \text{BELIEF} \rightarrow & (& 1 & 0 & \cdots) \end{array}$$

And just as before, we can associate each valued feature with a position in a vector of zeros and ones, then concatenate those rules into a matrix $\mathbf{P}$ (for projection).

(34) $$\mathbf{P} = \begin{array}{r} \\ \text{BELIEF} \\ \cdots \end{array} \begin{array}{ccc} \text{+TENSE} & \text{-TENSE} & \cdots \\ \end{array} \left(\begin{array}{ccc} 1 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{array} \right)$$

What does this buy us? It buys us labeling. To see this, we need to define how to get from a word's semantic components $\mathbf{S}$ to its syntactic distribution using the projection rules, encoded in $\mathbf{P}$. We'll state this directly and then walk through why this statement is reasonable. A word's distribution is given as a row in the matrix $\mathbf{D}$, which we define as the product of the semantic component matrix $\mathbf{S}$ and the projection matrix $\mathbf{P}$, given in (35a). If we view products as boolean $\wedge$ and sums as boolean $\vee$, this yields an expression in disjunctive normal form, as in (35b).

(35) a. $\mathbf{D} \equiv \mathbf{SP}$

b. $d_{vf} = \sum_k s_{vk} p_{kf} = \bigvee_k s_{vk} \wedge p_{kf}, \forall v \in \text{VERB}, \forall f \in \text{SYNTACTIC FEATURE}$
where $k \in \text{SEMANTIC COMPONENTS}$

So for instance, if v is *think* and f is [+TENSE], we know that d_{vf} will be (at least) one, since when k is BELIEF, both s_{vk} and p_{kf} are 1. We then say that, if d_{vf} is one, then verb i can take feature j. (One may wish to specify projection rules and syntactic distributions in terms of combinations of features, which in the limit yields subcategorization frames. This is easily done in the current setup by assuming that the columns of $\mathbf{P}$ and $\mathbf{D}$ are sets of valued features instead of single features.)

We submit that this setup gives us labeling for free. Since the rows of $\mathbf{P}$ are labeled and since the same row k of $\mathbf{P}$ is always multiplied with the same column k of $\mathbf{S}$. Thus, each column of $\mathbf{S}$ corresponds to a particular projection rule and, as a consequence, to the particular semantic component to which that projection rule corresponds. This is why we say that a top-down approach only considers a single L (or perhaps, has no notion of L at all).

But for the same reason that the top-down approach has no notion of L it also has no way of loosening the projection rules. As we discussed, this is problematic, since we need some way of capturing cross-linguistic differences in the way that semantic components like BELIEF are mapped onto the syntax. To see how this problem arises in the current setup, consider a very simple algorithm for reversing the projection rules: for each pairing of a verb v and a set of valued syntactic features—e.g., *think* S[+TENSE]—find a projection rule $\mathbf{p}_k$ that values those features the same way and flip the semantic component k on for verb v—i.e., set s_{vk} to one. But suppose that the projection matrix is the one given in (34). This is problematic for, e.g., Spanish learners, since they will receive examples like *want* S[+TENSE] and thus infer that *want* has a BELIEF component.

We are now in a position to state our proposal. Rather than encode $\mathbf{P}$ as in (34), we initialize it to one for every valuation of a feature that varies cross-linguistically. This yields a projection rule that can never match a particular structure, since by definition it contains conflicting

c. $$\begin{array}{rcccc} & & \text{TENSE} & \cdots & \cdots \\ \text{DESIRE} \rightarrow & (& -1 & 0 & \cdots) \end{array}$$

Why one would want this is unimportant for current purposes, but the intuition is roughly that projection rules do not appear to interact to cancel each other out. That is, having a particular semantic component only makes more structures possible—cf. *hope*, which has both BELIEF and DESIRE. Thus, under the current model, projection is *monotone nondecreasing*.

valuations of particular features. For instance, we posit the initial projection matrix in (36).

(36)
$$\mathbf{P} = \begin{array}{c|cccccccc} & +\text{TENSE} & -\text{TENSE} & +\text{COMP} & -\text{COMP} & +\text{IND} & -\text{IND} & \cdots \\ \hline \textsc{belief} & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

Part of the job of syntactic bootstrapping is then to filter out conflicting valuations until a concrete projection rule emerges, at which point the rule can in principle match actual feature valuations observed with a verb. The English learner must then perform the conversion in (37) and the Spanish learner must perform the conversion in (38).

(37)
$$\begin{array}{c|cccccccc} & +\text{TENSE} & -\text{TENSE} & +\text{COMP} & -\text{COMP} & +\text{IND} & -\text{IND} & \cdots \\ \hline \textsc{belief} & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$
$$\Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow$$
$$\begin{array}{c|cccccccc} & +\text{TENSE} & -\text{TENSE} & +\text{COMP} & -\text{COMP} & +\text{IND} & -\text{IND} & \cdots \\ \hline \textsc{belief} & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

(38)
$$\begin{array}{c|cccccccc} & +\text{TENSE} & -\text{TENSE} & +\text{COMP} & -\text{COMP} & +\text{IND} & -\text{IND} & \cdots \\ \hline \textsc{belief} & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$
$$\Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow$$
$$\begin{array}{c|cccccccc} & +\text{TENSE} & -\text{TENSE} & +\text{COMP} & -\text{COMP} & +\text{IND} & -\text{IND} & \cdots \\ \hline \textsc{belief} & 1 & 0 & 1 & 0 & 1 & 0 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

How do learners go about doing this conversion? This is where the notion of a featural anchor becomes important. We propose that the way learners perform the conversion in (37) and (38) is by setting the projection rule for BELIEF $\mathbf{p}_{\textsc{belief}}$ to the valuation they observed for the featural anchor corresponding to BELIEF—i.e., the valuation of their language's main clause.

We show how to implement this proposal algorithmically in the next section, but before moving on it is useful to see why going beyond the above formalization is necessary in the first place. One thing this formalization leaves open is how one should handle language internal irregularities in the projection rules. For instance, we noted in Section 3 that it is only a tendency for representationals to allow complementizer drop. But if our model values the projection rule for BELIEF based on a language's main clause, as in (39) for English, and if we require a perfect match with the projection rule, then the representationals that do not allow complementizer drop will not be labeled with the BELIEF component. This is even worse in a language like Spanish, where complementizer drop is not possible at all but where the declarative main clauses do not require complementizers.

(39)
$$\begin{array}{c|cccccccc} & +\text{TENSE} & -\text{TENSE} & +\text{COMP} & -\text{COMP} & +\text{IND} & -\text{IND} & \cdots \\ \hline \textsc{belief} & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$
$$\Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow$$
$$\begin{array}{c|cccccccc} & +\text{TENSE} & -\text{TENSE} & +\text{COMP} & -\text{COMP} & +\text{IND} & -\text{IND} & \cdots \\ \hline \textsc{belief} & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

This suggests that we need some way of allowing partial matches to the projection rules. In the next section, we implement this idea using a probabilistic model.

5 Implementing our proposal

In this section, we define a probabilistic model that implements the computational level description from the last section. We begin by describing the base model, which does not incorporate our proposal and is essentially a modified version of White's (2015, Ch. 3) model of syntactic bootstrapping. We then show how to implement our proposal by modifying this base model. We define an incremental learning algorithm that respects this model's assumptions and which we train on a dataset of syntactic distributions found in child-directed speech in Section 6 (Dudley et al., in prep).

5.1 Base model

Our model has two components, corresponding to two levels of abstraction. At the higher level—the *competence level*—the model describes the relationship between verbs' semantic features and their acceptability in frames with particular feature valuations. This is the level described in the formalization in the last section. As we did in that section, we call the representation of verbs' semantic features $\mathbf{S}$ (for semantic features), the representation of the relationship between semantic features and syntactic distribution $\mathbf{P}$ (for projection rules), and the representation of verbs' syntactic distributions $\mathbf{D}$.

At the lower level—the *performance level*—the model describes the relationship between verbs' acceptability in a particular syntactic context—what White (2015) refers to as its *competence distribution*—and the syntactic contexts it actually occurs in—which White refers to as its *performance distribution*. As is standard in the syntactic bootstrapping literature, we assume that the learner observes the performance distribution—or at least input from which it can be deterministically determined—and must infer the competence distribution.

5.1.1 The competence level

In the last section, we defined the competence level—the model's representation of verbs' semantic components $\mathbf{S}$ and projection rules $\mathbf{P}$—in terms of boolean values (zeros and ones). In this section, to allow for uncertainty about which semantic components a verb has and to model abstract projection rules, we define $\mathbf{S}$ and $\mathbf{P}$ in terms of probabilities. So, instead of the deterministic model in (40), we propose the probabilistic model in (41).

(40) **Deterministic model**
$\mathbf{S} \in \{0,1\}^{V\times K}$
$\mathbf{P} \in \{0,1\}^{K\times F}$
$\mathbf{D} \equiv \mathbf{SP} \in \{0,1\}^{V\times F}$

(41) **Probabilistic model**
$\mathbf{S} \in (0,1)^{V\times K}$
$\mathbf{P} \in (0,1)^{K\times F}$
$\mathbf{D} \equiv \mathbf{SP} \in (0,1)^{V\times F}$

So, instead of (42), we now have something of the form in (43). One way to think of this change is that $\mathbf{S}$ represents the model's certainty that a verb v does or does not have a particular semantic component k. If s_{vk} is close to one, then the model has high certainty that verb v has semantic component k, and if s_{vk} is close to zero, the model has high certainty that verb v does not have semantic component k.

(42)
$$\mathbf{S} = \begin{array}{c|ccc} & \text{BELIEF} & \text{DESIRE} & \cdots \\ \hline \text{think} & 1 & 0 & \cdots \\ \text{hope} & 1 & 1 & \cdots \\ \text{want} & 0 & 1 & \cdots \\ \cdots & \vdots & \vdots & \ddots \end{array}$$

(43)
$$\mathbf{S} = \begin{array}{c|ccc} & \text{BELIEF} & \text{DESIRE} & \cdots \\ \hline \text{think} & 0.95 & 0.20 & \cdots \\ \text{hope} & 0.89 & 0.86 & \cdots \\ \text{want} & 0.12 & 0.96 & \cdots \\ \cdots & \vdots & \vdots & \ddots \end{array}$$

Similarly, we can define the projection rules $\mathbf{P}$ probabilistically. Analogously to $\mathbf{S}$, $\mathbf{P}$ represents the model's certainty that a particular semantic component k does or does not project onto a particular syntactic feature f. If p_{kf} is close to one, then the model has high certainty that semantic component k does project onto syntactic feature f, and if p_{kf} is close to zero, the model has high certainty that semantic component k does not project onto syntactic feature f.

(44)
$$\mathbf{P} = \begin{array}{c|ccccccc} & \text{+TENSE} & \text{-TENSE} & \text{+COMP} & \text{-COMP} & \text{+IND} & \text{-IND} & \cdots \\ \hline \text{BELIEF} & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

(45)
$$\mathbf{P} = \begin{array}{c|ccccccc} & \text{+TENSE} & \text{-TENSE} & \text{+COMP} & \text{-COMP} & \text{+IND} & \text{-IND} & \cdots \\ \hline \text{BELIEF} & 0.96 & 0.03 & 0.06 & 0.92 & 0.14 & 0.02 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

The final component of the competence level, the syntactic distributions $\mathbf{D}$, is defined in terms of $\mathbf{S}$ and $\mathbf{P}$. When $\mathbf{S}$ and $\mathbf{P}$ are boolean-valued, a natural way of defining the $\mathbf{D}$ is in terms of disjunctive normal form. We can naturally extend this idea to probabilities using products, the analogy of boolean $\wedge$, and maximums, the analogy of boolean $\vee$.[7]

(46) $d_{vf} \equiv \bigvee_k s_{vk} \wedge p_{kf}$

(47) $d_{vf} \equiv \max\{s_{vk}p_{kf} \mid k \in \text{SEMANTIC COMPONENTS}\}$

Under the definition in (47), a verb's syntactic distribution is determined by asking, for each semantic components, what the probability is (i) that the verb has that semantic component and (ii) that that semantic component projects onto a particular feature. The probability that the verb takes a particular feature is then determined by the largest such probability.

This probabilistic model is useful for the model has little data to go on, but as the learner gathers more information about a verb's distribution, we would like it to converge toward a set of semantic components it believes each verb has. To do this, we aim to induce *sparsity* in the semantic component and projection rule representations as the model gets more data. Inducing sparsity amounts to encouraging the model to set the entries of $\mathbf{S}$ and $\mathbf{P}$ close to either one or zero. Thus, as the model gains more data, it should begin to look more and more like the strict boolean version.

To induce sparsity, we place independent symmetric beta distributions on the entries of $\mathbf{S}$ and $\mathbf{P}$. These beta distributions are uniform—Beta(1,1)—to begin with, not preferring any particular value in (0,1), but over the course of learning, they become sparser, preferring values

[7] This extension is a direct application of the fuzzy logic notion of a *t-norm* and *t-conorm* (Zadeh, 1965; Klement et al., 2013). The specific t-norm we use here is known as the *product t-norm*. See also work in Probabilistic Soft Logic (Kimmig et al., 2012).

closer to zero or one, as the parameters of the distribution are tuned toward zero.[8]

5.1.2 The performance level

To complete the base model, we need some way of linking the competence level to the observed data. We do this very straightforwardly by assuming that each datapoint consisting of verb v and syntactic feature combination $\mathbf{x}$ is generated by sampling the value x_f for syntactic feature f from a Bernoulli distribution. The likelihood for performance level is then given in (48).[9]

(48) $\mathbb{P}(v, \mathbf{x} \mid \mathbf{d}_v) = \prod_f \text{Bernoulli}(x_f; d_{vf}) = \prod_f d_{vf}^{x_f}(1 - d_{vf})^{1-x_f}$

In this way, the likelihood function for our model is very similar to Alishahi and Stevenson's (2008), though our representation of the competence distributions and incorporation of labeled projection rules differs substantially from their model.[10] The full model has a close resemblance to a Restricted Boltzmann Machine or Harmonium (Smolensky, 1986).

5.2 Incorporating featural anchors

Incorporating featural anchors into the base model is straightforward, though it cannot be done in quite the same manner described above, wherein the projection rule is valued directly by the relevant featural anchor. If it were valued this way, we would not gain the benefits of partial matching that moving to a probabilistic model promises, since setting parts of the projection rule to one means that a particular syntactic context would need to match the projection rule perfectly. We noted that this was a problem for verbs that do not allow complementizer drop.

One way to remedy this is to treat main clauses as the same sort of object as subordinate clauses—i.e., as sets of valued syntactic features associated with some special verb. If main clauses are assimilated with subordinate clauses in this way, we allow the model to retain the sort of uncertainty necessary for allowing partial matches, since whatever special verb main clauses are associated with will—if the model works correctly—share a semantic component with verbs like *think*.

The idea that main clauses are in fact subordinate clauses of a particular kind of verb (or set of verbs) is an old idea instantiated most famously by Ross's (1970) Performative Hypothesis (see also Rizzi, 1997; Ambar, 1999; Krifka, 2001; Ginzburg and Sag, 2001; Speas and Tenny, 2004; Hacquard, 2010). Under the Performative Hypothesis, all main clauses are actually embedded under special null verbal elements corresponding to the sorts of discourse moves defined by Austin (1975) (see also Urmson, 1952). Following Hacquard (2010), and others, we call the special element under which declarative main clauses are embedded ASSERT.

(Importantly, note that one need not accept the Performative Hypothesis as a hypothesis about syntactic representation in accepting our model. The Performative Hypothesis merely provides a clean way of implementing our model, which could be implemented in various other ways.)

[8] For the current model, this is done globally by tuning the beta hyperparameters according to the total number of datapoints seen. An alternative model might tune these beta hyperparameters in a verb-specific way according to the amount of data seen for that particular verb.

[9] One thing that we do not model here is how the verb v is chosen. This could easily be incorporated into this model by including a term for the prior probability of v, though we leave it out for simplicity. (See White 2015, where such a model is discussed extensively.)

[10] Many other models work in terms of raw or modified count distributions (Schulte im Walde, 2000; Schulte im Walde and Brew, 2002; Schulte im Walde, 2003, 2006; Korhonen, 2002; Buttery and Korhonen, 2005; Buttery, 2006). This includes White's (2015) on which the current model is based. One potentially useful thing about directly modeling counts is that it allows a model to control for power law distributions by using particular kinds of likelihood functions, such as those founded on negative binomial distributions (Church and Gale, 1995) or Bayesian nonparametrics (Goldwater et al., 2011). See Piantadosi 2014 for a review of power law distribution models.

To encode the fact that BELIEF corresponds to a particular component, we initialize the model in such a way that ASSERT has only a single semantic component (and no other semantic components) with total certainty. We then stipulate that the semantic component ASSERT certainly has is the BELIEF component, and we disallow the model from changing ASSERT's semantic components.

(49) $$\mathbf{S} = \begin{array}{l|ccc|} & \text{BELIEF} & \cdots & \cdots \\ \text{ASSERT} & 1.00 & 0.00 & \cdots \\ \text{think} & 0.95 & 0.20 & \cdots \\ \text{hope} & 0.89 & 0.86 & \cdots \\ \text{want} & 0.12 & 0.96 & \cdots \\ \cdots & \vdots & \vdots & \ddots \end{array}$$

This effectively forces the model to explain main clause features using only a single projection rule—the one corresponding to the semantic component ASSERT certainly has—but it also allows the model to have partial matches, since once it posits that other verbs have this semantic component, the probability of other feature values—e.g., [+COMP]—increase.

This shows that our proposal requires only very minimal additions to current models of syntactic bootstrapping. Indeed, using a similar strategy, our model can also be incorporated into Alishahi and Stevenson's model (though such an addition would run somewhat counter to the constructionist (and constructivist) position that Alishahi and Stevenson take).

5.3 Learning algorithms

Many different kinds of learners can be defined to respect this model's assumptions to varying degrees. Here, we define an incremental learner, which observes pairings of verbs and syntactic features one at a time and makes inferences after each observation. This learner is implemented using a form of stochastic projected gradient descent with adaptive gradient. We do not delve into the specifics of this algorithm, though we do give a high-level description of what it is doing.

The learner begins with randomly initialized matrices $\mathbf{S}$ and $\mathbf{P}$ of values in (0,1). Upon receiving a particular datapoint consisting of a verb v and a syntactic structure represented as a boolean string $\mathbf{x}$, the learner calculates how likely that datapoint is given the current model using the Bernoulli likelihood function given above and repeated in (50).

(50) $$\mathbb{P}(v, \mathbf{x} \mid \mathbf{d}_v) = \prod_f d_{vf}^{x_f}(1 - d_{vf})^{1-x_f} = \prod_f \max_k\{s_{vk}p_{kf}\}^{x_f}(1 - \max_k\{s_{vk}p_{kf}\})^{1-x_f}$$

The learner then attempts to change the semantic representation for the verb $\mathbf{s}_v$ and the projection rules $\mathbf{P}$ so that they give a higher likelihood to the data—but not so much that the current level of sparsity (closeness to zero or one) induced by the priors is violated. The adaptive gradient piece of the learner ensures that the changes to the verb's semantic representation are not very extreme if the verb has been seen many times before but are potentially extreme if the verb is very infrequent.

6 Experiment

We now apply our learning algorithm to a dataset of English child-directed speech. For the purposes of this experiment, we assume two semantic components: one associated with ASSERT (BELIEF) and the rest not associated with any particular semantic component. For a full model of syntactic bootstrapping, more components would be necessary. For instance, White (2015) and White et al. (under reviewa) show that the syntax carries quantifiable information about propositional attitude verb semantic components like FACTIVITY, COMMUNICATIV-

ITY, PERCEPTION, etc. But since this experiment serves more as a proof of concept and since it makes the interpretation easier, we do not incorporate more components into the current model. Furthermore, we only analyze the component corresponding to belief, since this is the only one our model actually labels.

We begin by describing the dataset along with some necessary preliminaries on correcting for sampling bias in this dataset. We then show the results of fitting our algorithm to these corrected data.

6.1 Data

Dudley et al. (in prep), who investigate the cues to factivity available in the syntactic (performance) distributions found in child-directed speech, annotate the syntactic features associated with 77 clause embedding predicates found in the Gleason corpus (Masur and Gleason, 1980) in CHILDES (MacWhinney, 2014b,a). These predicates are shown in (51).

(51) be able, be afraid, agree, be anxious, approve, ask, assure, be aware, believe, bet, care, be certain, complain, consider, be curious, decide, discover, doubt, be eager, enjoy, expect, explain, figure, find, forget, gather, be glad, guess, be happy, hate, hear, hope, ignore, imagine, be important, be impossible, interest, be interesting, judge, know, learn, like, love, manage, mean, need, be nervous, notice, order, predict, prefer, pretend, propose, be proud, prove, realize, recall, regard, remember, remind, say, see, seem, show, be sorry, suppose, be sure, tell, think, try, understand, want, be willing, wish, wonder, be worth

For each of the predicates with less than 100 occurrences, Dudley et al. annotate all occurrences; and for each predicate with 100 or more occurrences, they annotate a randomly selected 100 occurrences. Each occurrence is annotated for a wide variety of features, including features of the predicate's subject as well as non-subordinate clause complements. Here, we focus only on features of the subordinate clause, listed in (52).

(52) a. [+/- EMBEDDED SUBJECT]
b. [+/- COMPLEMENTIZER]
c. [+/- TENSE]
d. [+/- INFINITIVAL]

This means that each observation is constituted by a verb paired with a string of eight boolean values—one for each valuation of the feature. (As a reminder, the reason we need distinct boolean values for each feature valuation is that some verbs—e.g., *hope*, *remember*, *decide*, etc.—can occur in subcategorization frames with different valuations for a features—e.g., *hope that*, which is [+ TENSE], and *hope to*, which is [- TENSE].)

6.2 Resampling procedure

We noted above that in Dudley et al.'s data the proportion of datapoints for each verb does not match its proportion in child-directed speech. This means that lower frequency verbs are overrepresented in this dataset. We correct for this overrepresentation using a stratified resampling procedure. The basic idea behind this procedure is that we need to replicate the sort of evidence a learner might actually observe, and so we need to make sure the dataset is representative of the true verb frequencies.

To do this, we first count the total number of tokens in CHILDES for each of the 77 verbs in the dataset, "hallucinating" a count for ASSERT equal to half the sum of the frequencies for these 77 verbs. We then find the frequency of each verb relative to the 77 others plus ASSERT. Using this relative frequency, we draw 100 samples from a multinomial distribution (n=25000)

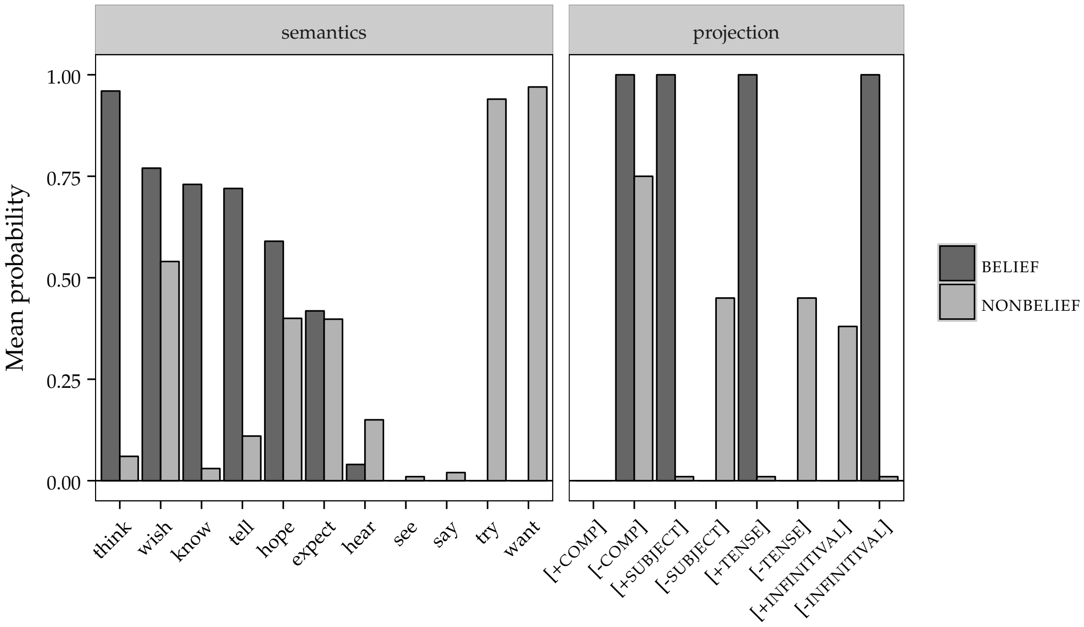

Figure 1: Mean probability of BELIEF and NONBELIEF semantic components over 100 runs

representing the number of times each verb will be seen in each resampled dataset. For each sample, for each verb, datapoints are then drawn uniformly without replacement until that verb's count for that sample is reached.

6.3 Fitting

We run our algorithm on each of these resampled datasets described above—reinitializing the algorithm between each run and randomly permuting the order in which the algorithm observes datapoints in that dataset. Over the course of learning, we gradually increase the sparsity of the semantic representations and the projection rules by decreasing the parameters of the beta prior from α=β=1 to 0.5 using a function very similar to search-then-converge (Darken and Moody, 1990). In practice, this is sufficient to induce the algorithm to use only two values in the semantic representations and projection rules—one very close to zero and one very close to one—and the model found such a solution on all trials for all verbs. For the purposes of reporting the results, we discretize these values to zero and one, respectively, and report the proportion of times over the 100 runs that this value was one.

6.4 Results

Figure 1 shows the mean probability of the BELIEF and NONBELIEF semantic components over the 100 runs of our algorithm. We see that on virtually all runs, the model learns that *think* has a BELIEF component and that *want* does not. Further, as we expect, the BELIEF component projects onto exactly the main clause features, and no others.

Further, the algorithm converges to this solution for *think* and *want* relatively quickly. Figure 2 shows the mean probability of the BELIEF and NONBELIEF semantic components over the 100 runs of our algorithm as a function of the number of datapoints the algorithm has seen—i.e., not just the number of observations of *think* or *want* alone. We see that, by about 1500 observations of the 25000 total observations, the algorithm has learned (i) that *want* does not have a BELIEF component but does have a NONBELIEF component and (ii) that *think* does not have a NONBELIEF component (at least when only two components are considered).

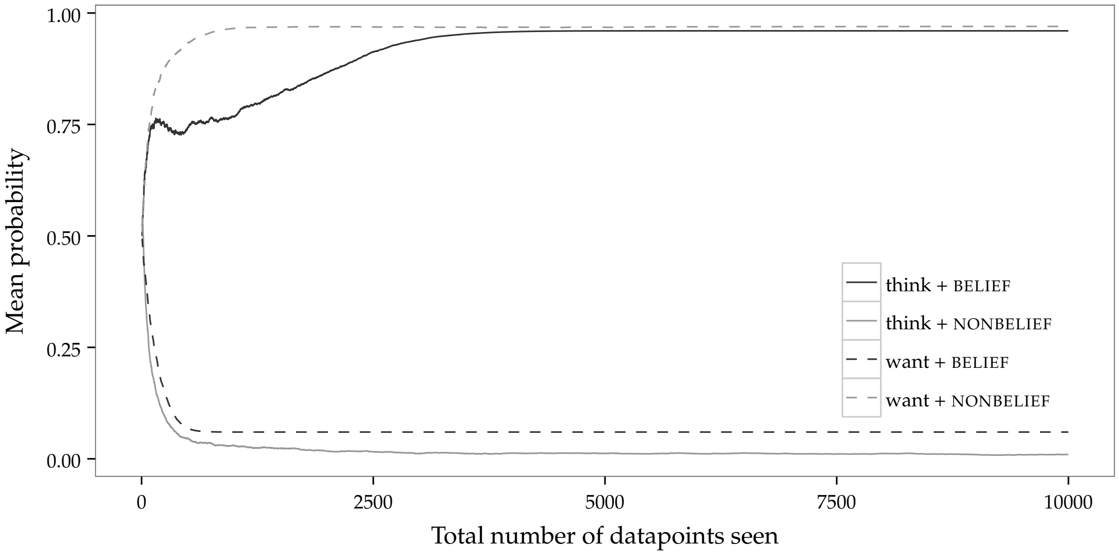

Figure 2: Mean probability of BELIEF and NONBELIEF semantic components for *think* and *want* over 100 runs as a function of number of total datapoints seen

Interestingly, the model takes longer (on average) to ascertain that *think* has a BELIEF component. In looking at individual runs of the algorithm, this appears to arise from two phenomena. First, *think* is seen half as often as *want*, and so it is seen less and thus updated less often. This frequency difference, however, does not explain the differential speed with which the model approaches certainty that *think* does not have a NONBELIEF component. And so, second, this differential appears to be explained by the fact that the change in the probability of a BELIEF component for *think* shows higher sustained variability than that for the NONBELIEF component. This can be seen in Figure 3 and suggests that the learner has more uncertainty about whether *think* has a BELIEF component for longer. This is perhaps unsurprising, since *think*, and most other verbs, take more than just a single complement type. For instance, one fairly common complement type found with *think* is the propositional anaphor *so*.

(53) I think so.

But though *so* in this case likely refers to the same type of thing as a tensed complement—i.e., a proposition—our model cannot take advantage of this, since (i) it was not coded for, and (ii) even if it were, it would not share syntactic features with a declarative main clause. We return to the question of how a learner might use information from an anaphor like *so* in Section 7.

Turning now to the other verbs in Figure 1, we see that our algorithm does remarkably well in labeling the correct verbs with a BELIEF component. For instance, *know*, *hope*, *wish*, and *tell* are all labeled with a BELIEF component more than half the time. Further, verbs like *hope* and *wish*, which also have DESIRE components are labeled with a NONBELIEF component fairly often. This is quite surprising, since these verbs are seen very rarely on average: *hope* and *wish* are only seen 48 and 31 times on average, respectively, out of the 25000 datapoints the model sees total. Compare these numbers to the number of times *think* (1337) and *want* (2889) are seen on average.

Further, the behavior for *hope* in particular aligns well with recent experimental findings. In an experiment that makes both beliefs and desires salient, Harrigan (2015) and Harrigan et al. (2016) show that three year old English learners tend to attribute to *hope* a BELIEF meaning when it appears with a tensed subordinate clause but a DESIRE meaning when it appears with an infinitival subordinate clause. This suggests that their semantic representation of the verb is not fully stable—their interpretation being highly influenced by syntactic context.

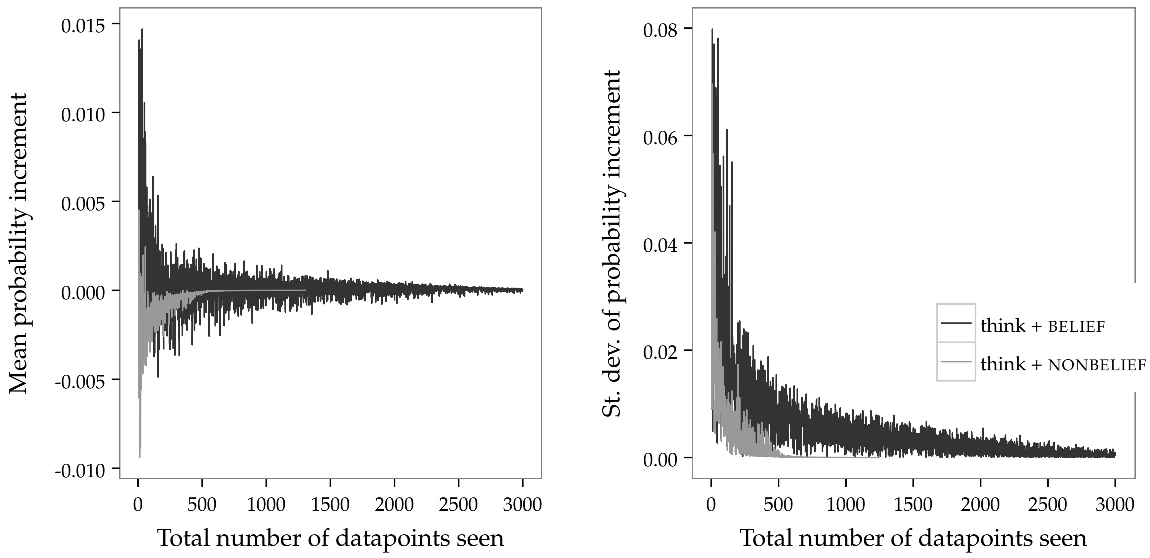

Figure 3: Mean and standard deviation of probability increment for *think* over 100 runs as a function of number of total datapoints seen

Other verbs do not fair as well, including *say* and the two most frequent perception verbs *see* and *hear*. The verbs do not appear to be associated with either semantic component for two reasons. First, *say* takes many quotes of varying levels of syntactic complexity—from one word quotes to phrasal quotes to fully clausal quotes—and all three take noun phrase complements fairly often.[11] These structures are not captured by the projection rules in this experiment. This actually may be a problem more generally for a main clause, and it may require us to posit something that weights the clausal complements that are observed more highly in the likelihood function. This approach seems reasonable given the fact that noun phrase complements are not particularly informative in this domain (White et al., under reviewa), though ideally we would like the model to learn the correct weighting.

Second, and pertaining mostly to the perception verbs, the fact that small clause complements, such as (54), are untensed may push the model away from viewing perception verbs as having a BELIEF component.

(54) You should have heard her sing in the car last night.

(55) I heard that she sang in the car last night.

One possibility with regard to the perception verbs is that it is only learned later that (some of) these verbs have a belief component. Indeed, the *hear* in (55) may not even be the same *hear* as in (54).

6.5 Discussion

In this experiment, we applied the learning algorithm we define in Section 5 to Dudley et al.'s (in prep) dataset of syntactic distributions from child-directed speech. We showed that this algorithm does remarkably well in labeling the correct verbs with a BELIEF component, with only a few (interesting) exceptions.

[11] The reason *say* takes so many quotes likely arises from a bias in the corpus. Part of the corpus contains transcripts of book reading.

7 General discussion

In this paper, we propose a novel solution to the labeling problem in syntactic bootstrapping that is a hybrid of the top-down and bottom-up approaches. Under this proposal, we retain the framework posited by the top-down approach, which solves the labeling problem by positing innate projection rules, but we allow these rules to be abstract in a highly constrained way.

We motivate this proposal by noting that neither the top-down nor the bottom-up approaches solve the labeling problem for belief and desire verbs. The top-down approach is brittle in the face of cross-linguistic variability, while the bottom-up approach makes unrealistic assumptions about the data learners have access to.

We show that our proposed solution can deal with the labeling problem given theoretically justified featural anchors for particular labels, and using belief and desire predicates as a case study, we implement a computational model that incorporates the labeling mechanism we propose. Finally, we provide a proof of concept fit of this model to data derived from child-directed speech.

In the remainder of the paper, we discuss the implications of this work for the theory of verb learning and its relationship to linguistic theory more generally.

7.1 The explanatory power of featural anchors

The main substantitve addition we make in this paper to the theory of syntactic bootstrapping is the notion of a featural anchor, which is itself a class of structures known prior to the selection of a projection rule. One question this addition raises is to what extent learners can easily discover the featural anchor itself. A deflationary response to our proposal might argue that we have merely pushed the job of verb learning back to discovering which classes of structures constitute featural anchors; isn't the job of figuring out which syntactic features are indicative of a particular anchor just as hard as learning a verb itself?

Yes and no. Yes, because it is true that, at the end of the day, one must identify some structure as a declarative main clause, and this must presumably be done by perceiving that, in using a particular structure, an utterer of a declarative main clause intends the utterance to be taken as part of a particular conversational move—such as an assertion (Austin, 1975; Stalnaker, 1978). That is, the learner must be able to identify the illocutionary force intended for the utterance.

No, because illocutionary force is a concept that is presumably prerequisite to learning a language in a first place.[12] Indeed, children appear to be adept at recognizing an utterance's illocutionary force quite early (Spekman and Roth, 1985). This is to say that, though language is clearly not *for* communication, the data a learner uses to learn their language tends to come wrapped in communicative acts, which the learner presumably has no problem perceiving as such.

And no, because the mapping between illocutionary force and syntactic structure is relatively stable within a language: assertions, at least as conveyed by clauses, are conveyed by clauses with the same syntactic features—in English, [+TENSE, -COMP, . . .]. And insofar as an assertion is not conveyed by a clause—such as when it is conveyed by a polarity particle or a fragment—we submit that, if a learner has enough syntactic knowledge to represent a clause as a set of features, they have enough syntactic knowledge to represent that the valuation of those features is dependent on the fact that that clause is a syntactic object in the first place.

This question of recognizing a syntactic class of complements, such as a clause, is related to the issue we saw our algorithm having with nominal and propositional anaphor complements, like *so*. These complements cannot be valued for the same sorts of features that a clause can be—indeed, they don't appear to be valued as such at any level of syntactic representation

[12] Thanks go to Alexander Williams for driving this point home to us.

(Hankamer and Sag, 1976)—but as it stands our model views them as simply unvalued. But rather than view them as unvalued for these features, it seems that the distinction in syntactic class must be baked into the model itself. That is, the syntactic features the model pays attention to in making an inference from a particular piece of data must be dependent on the syntactic class involved in that datum. In the abstract, we need to incorporate some decision tree-like representation into the model. One way this might be implemented is by employing a likelihood function that incorporates a hurdle model. (See White et al. under reviewb for a recent use of hurdle models in a related domain.)

Beyond providing a case for a decision tree-like structure for syntactic feature valuation, the case of propositional anaphors may also suggest that the mechanisms used by syntactic bootstrapping to infer a verb's meaning may need to incorporate a notion of semantic type over and above that given by syntactic type. (See White and Rawlins in prep for evidence that semantic type signatures can be extracted from syntactic distribution.)

7.2 Generalizing our argument

The analysis of the belief-desire distinction that we give in Section 3 takes the following general form.

1. the semantic class of verbs C—e.g., belief verbs—in language L tend to embed structures—e.g., subordinate clauses—with features F_L—e.g., TENSE, V2, etc.
2. features F_L are all valued in structures X_L, which is an independently motivated class of structures found in all languages $L \in \mathcal{L}$
3. F_L for all $L \in \mathcal{L}$ varies on only features $F_{\mathcal{L}}$
4. this suggests
   (a) an abstract projection rule $C \rightarrow [+/\text{-}f : \forall f \in F_{\mathcal{L}}]$
   (b) a featural anchor X_L for each language L

One question that arises is to what extent there are other abstracted projection rules and featural anchors. There seems to us to be at least two potential candidates: preferential verbs, like *want*, and factive and interrogative verbs like *know* and *wonder*.

With respect to preferential verbs, as we noted in Section 3, preferential verbs seem to have a cross-linguistically variable distribution. For instance, one could not label preferentials by assuming they were tenseless, subjectless, or lacked a complementizer, as they do in English, since, e.g., in Spanish, all of these things are present (assuming that the Spanish learner assumes a covert pronoun in the case of *pro*-drop).

(56) a. I want to go home.
b. Quiero que vayas a la casa.
want.1S.PRES that Peter go.PRES.SBJ to the house.

This may suggest that one should expect a featural anchor for the projection rule labeled with the DESIRE component. One plausible such anchor are imperative clauses. Note that in both English and Spanish, imperatives share features with the sorts of subordinate clauses found under preferential verbs.[13]

(57) a. Go home.

[13] One difficulty that may arise here, however, is that the mapping between subjunctive and preferentiality is not necessarily as robust as that between, e.g., indicative and representationality.

b. Vayas a la casa.
go.2S.PRES.SBJ to the house.

This may be related to the fact that preferential verbs can be used to make indirect requests, as we saw in Section 3.

(58) I want you to go to your room!

With respect to factive and interrogative verbs, Dudley et al. (in prep) show that a large part of children's experience with *know* is in the frame *do you know Q?* We would like to suggest that, here again, the child may be able exploit the syntactic parallels between direct and indirect speech acts: *know* is used to ask indirect questions. Children might infer from this parallel that the meaning of *know* must be one that relates the subject to the answer of that question.

(59) Do you know where the keys are?

These two additional examples may suggest that learners are able to exploit the relationship between various kinds of main clause syntax—declarative, imperative, and interrogative—and the syntax of subordinate clauses more generally. In future work, we aim to investigate this possibility.

References

Alishahi, Afra, and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science* 32:789–834.

Ambar, Manuela. 1999. Aspects of the syntax of focus in Portuguese. *The grammar of focus* 24:23.

Anand, Pranav, and Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics* 6:1–59.

Anand, Pranav, and Valentine Hacquard. 2014. Factivity, belief and discourse. In *The Art and Craft of Semantics: A Festschrift for Irene Heim*, ed. Luka Crnič and Uli Sauerland, volume 1, 69–90. Cambride, MA: MIT Working Papers in Linguistics.

Austin, John Langshaw. 1975. *How to do things with words*. Oxford university press.

Baillargeon, Renée, Rose M. Scott, and Zijing He. 2010. False-belief understanding in infants. *Trends in cognitive sciences* 14:110–118.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2012. Modeling the acquisition of mental state verbs. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, 1–10. Association for Computational Linguistics.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2013. Acquisition of Desires before Beliefs: A Computational Investigation. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL)*.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2014a. Gradual Acquisition of Mental State Meaning: A Computational Investigation. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2014b. Learning verb classes in an incremental model. *ACL 2014* 37.

Bolinger, Dwight. 1968. Postposed main phrases: an English rule for the Romance subjunctive. *Canadian Journal of Linguistics* 14:3–30.

Buttery, Paula. 2006. Computational models for first language acquisition. Doctoral Dissertation, University of Cambridge.

Buttery, Paula, and Anna Korhonen. 2005. Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Carter, Richard. 1976. Some linking regularities. *On Linking: Papers by Richard Carter Cambridge MA: Center for Cognitive Science, MIT (Lexicon Project Working Papers No. 25)* .

Chomsky, Noam. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Walter de Gruyter.

Church, Kenneth W., and William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering* 1:163–190.

Connor, Michael, Cynthia Fisher, and Dan Roth. 2013. Starting from scratch in semantic role labeling: Early indirect supervision. In *Cognitive aspects of computational language acquisition*, 257–296. Springer.

Darken, Christian, and John Moody. 1990. Note on learning rate schedules for stochastic optimization. In *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*, NIPS-3, 832–838. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Dayal, Veneeta, and Jane Grimshaw. 2009. Subordination at the interface: the Quasi-Subordination Hypothesis.

De Villiers, Jill G., and Peter A. De Villiers. 2000. Linguistic determinism and the understanding of false belief. *Children's Reasoning and the Mind* 191–228.

De Villiers, Jill G., and Jennie E. Pyers. 2002. Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development* 17:1037–1060.

Dudley, Rachel, Meredith Rowe, Valentine Hacquard, and Jeffrey Lidz. in prep. On the syntactic distribution of mental state verbs in child-directed speech.

Farkas, Donka. 1985. *Intensional descriptions and the Romance subjunctive mood*. Taylor & Francis.

Frank, Michael C., Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20:578–585.

Giannakidou, Anastasia. 1997. The landscape of polarity items. Doctoral Dissertation, University of Groningen.

Gillette, Jane, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition* 73:135–176.

Ginzburg, Jonathan, and Ivan Sag. 2001. *Interrogative investigations*. Stanford: CSLI publications.

Giorgi, Alessandra, and Fabio Pianesi. 1997. *Tense and Aspect: Form Semantics to Morphosyntax*. Oxford: Oxford University Press.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language acquisition* 1:3–55.

Gleitman, Lila R., Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. 2005. Hard words. *Language Learning and Development* 1:23–64.

Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *The Journal of Machine Learning Research* 12:2335–2382.

Grimshaw, Jane. 1981. Form, function and the language acquisition device. In *The Logical Problem of Language Acquisition*, 165–182.

Grimshaw, Jane. 1990. *Argument structure*. Cambridge, MA: MIT Press.

Grimshaw, Jane. 1994. Lexical reconciliation. *Lingua* 92:411–430.

Grimshaw, Jane. 2009. That's nothing: the grammar of complementizer omission.

Gruber, Jeffrey Steven. 1965. Studies in lexical relations. Doctoral Dissertation, Massachusetts Institute of Technology.

Hacquard, Valentine. 2010. On the event relativity of modal auxiliaries. *Natural Language Semantics* 18:79–114.

Hacquard, Valentine. 2014. Bootstrapping attitudes. In *Semantics and Linguistic Theory*, volume 24, 330–352.

Hale, Ken, and Samuel Jay Keyser. 2002. *Prolegomena to a Theory of Argument Structure*. Cambridge, MA: MIT Press.

Hankamer, Jorge, and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic Inquiry* 391–428.

Harrigan, Kaitlyn. 2015. Syntactic bootstrapping in the acquisition of attitude verbs. Doctoral Dissertation, University of Maryland.

Harrigan, Kaitlyn, Valentine Hacquard, and Jeffrey Lidz. 2016. Syntactic Bootstrapping in the Acquisition of Attitude Verbs: think, want and hope. In *Proceedings of WCCFL 33*, ed. Kyeong-min Kim, Pocholo Umbal, Trevor Block, Queenie Chan, Tanie Cheng, Kelli Finney, Mara Katz, Sophie Nickel-Thompson, and Lisa Shorten. Cascadilla Proceedings Project.

Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of semantics* 9:183–221.

Hooper, Joan B. 1975. On Assertive Predicates. In *Syntax and Semantics*, ed. John P. Kimball, volume 4, 91–124. New York: Academy Press.

Kako, Edward. 1997. Subcategorization Semantics and the Naturalness of Verb-Frame Pairings. *University of Pennsylvania Working Papers in Linguistics* 4:11.

Kimmig, Angelika, Stephen Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 1–4.

Klement, Erich Peter, Radko Mesiar, and Endre Pap. 2013. *Triangular norms*, volume 8. Springer Science & Business Media.

Korhonen, Anna. 2002. Subcategorization Acquisition. Doctoral Dissertation, University of Cambridge.

Krifka, Manfred. 2001. Quantifying into question acts. *Natural Language Semantics* 9:1–40.

Landau, Barbara, and Lila R. Gleitman. 1985. *Language and experience: Evidence from the blind child*, volume 8. Harvard University Press.

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

Lewis, Shevaun. 2013. Pragmatic enrichment in language processing and development. Doctoral Dissertation, University of Maryland.

Lidz, Jeffrey, Henry Gleitman, and Lila Gleitman. 2004. Kidz in the 'hood: Syntactic bootstrapping and the mental lexicon. In *Weaving a Lexicon*, ed. D.G. Hall and S.R. Waxman, 603–636. Cambridge, MA: MIT Press.

MacWhinney, Brian. 2014a. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

MacWhinney, Brian. 2014b. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Marr, David. 1982. Vision: a computational investigation into the human representation and processing of visual information. *Henry Holt and Co.* .

Masur, Elise F., and Jean B. Gleason. 1980. Parent–child interaction and the acquisition of lexical information during play. *Developmental Psychology* 16:404.

Medina, Tamara Nicol, Jesse Snedeker, John C. Trueswell, and Lila R. Gleitman. 2011. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences* 108:9014–9019.

Onishi, Kristine H., and Renée Baillargeon. 2005. Do 15-month-old infants understand false beliefs? *Science* 308:255–258.

Papafragou, Anna, Kimberly Cassidy, and Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition* 105:125–165.

Perner, Josef, Manuel Sprung, Petra Zauner, and Hubert Haider. 2003. Want That is Understood Well before Say that, Think That, and False Belief: A Test of de Villiers's Linguistic Determinism on German–Speaking Children. *Child development* 74:179–188.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review* 21:1112–1130.

Pinker, Steven. 1984. *Language learnability and language development*. Harvard University Press.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pinker, Steven. 1994. How could a child use verb syntax to learn verb semantics? *Lingua* 92:377–410.

Portner, Paul. 1992. Situation theory and the semantics of propositional expressions. Doctoral Dissertation, University of Massachusetts, Amherst.

Portner, Paul, and Aynat Rubinstein. 2013. Mood and contextual commitment. In *Semantics and Linguistic Theory*, volume 22, 461–487.

Quer, Josep. 1998. Mood at the Interface. Doctoral Dissertation, Utrecht Institute of Linguistics, OTS.

Rizzi, Luigi. 1997. The fine structure of the left periphery. In *Elements of grammar*, 281–337. Springer.

Ross, John R. 1970. On declarative sentences. *Readings in English transformational grammar* 222:272.

Ross, John Robert. 1973. Slifting. In *The formal analysis of natural languages*, ed. Maurice Gross, Morris Halle, and Marcel-Paul Schützenberger, 133–169. The Hague: Mouton de Gruyter.

Scheffler, Tatjana. 2009. Evidentiality and German attitude verbs. *University of Pennsylvania Working Papers in Linguistics* 15.

Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117:1034–1056.

Smith, Linda, and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106:1558–1568.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, ed. James McClelland and David Rumelhart, volume 1. Cambridge, MA: MIT Press.

Snedeker, Jesse, and Lila Gleitman. 2004. Why it is hard to label our concepts. In *Weaving a Lexicon*, 257–294. Cambridge, MA: MIT Press.

Speas, Peggy, and Carol Tenny. 2004. Configurational properties of point of view roles. *Asymmetry in grammar* 1:315–345.

Spekman, Nancy J., and Froma P. Roth. 1985. Preschool children's comprehension and production of directive forms. *Journal of Psycholinguistic Research* 14:331–349.

Stalnaker, Robert. 1978. Assertion. *Syntax and Semantics (New York Academic Press)* 9:315–332.

Stalnaker, Robert. 1984. *Inquiry*. Cambridge University Press.

Stone, Marshall H. 1936. The theory of representation for Boolean algebras. *Transactions of the American Mathematical Society* 40:37–111.

Truckenbrodt, Hubert. 2006. On the semantic motivation of syntactic verb movement to C in German. *Theoretical Linguistics* 32:257–306.

Trueswell, John C., Tamara Nicol Medina, Alon Hafri, and Lila R. Gleitman. 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology* 66:126–156.

Urmson, James O. 1952. Parenthetical verbs. *Mind* 61:480–496.

Villalta, Elisabeth. 2000. Spanish subjunctive clauses require ordered alternatives. In *Semantics and Linguistic Theory*, volume 10, 239–256.

Villalta, Elisabeth. 2008. Mood and gradability: an investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31:467–522.

de Villiers, Jill G. 2005. Can Language Acquisition Give Children a Point of View? In *Why language matters for theory of mind*, ed. Janet W. Astington and Jodie A. Baird, 186–219.

Schulte im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 747–753. Association for Computational Linguistics.

Schulte im Walde, Sabine. 2003. Experiments on the Automatic Induction of German Semantic Verb Classes. Doctoral Dissertation, Universität Stuttgart.

Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32:159–194.

Schulte im Walde, Sabine, and Chris Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 223–230. Association for Computational Linguistics.

White, Aaron Steven. 2015. Information and incrementality in syntactic bootstrapping. Doctoral Dissertation, University of Maryland.

White, Aaron Steven, Valentine Hacquard, and Jeffrey Lidz. under reviewa. Projecting attitudes.

White, Aaron Steven, and Kyle Rawlins. in prep. A computational model of S-selection. In *Semantics and Linguistic Theory*, volume 26.

White, Aaron Steven, Drew Reisinger, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. under reviewb. *Computational linking theory*.

Yu, Chen, and Linda B. Smith. 2007. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science* 18:414–420.

Yu, Chen, and Linda B. Smith. 2012. Modeling cross-situational word–referent learning: Prior questions. *Psychological review* 119:21.

Zadeh, Lotfi A. 1965. Fuzzy sets. *Information and control* 8:338–353.