Setting Syntactic Parameters with Implicit Negative Evidence: The Case of Zero-derived Causatives in English

Isaac Gould
Department of Linguistics and Philosophy
Massachusetts Institute of Technology
`igould@mit.edu`

Abstract

In this paper, I introduce a learning challenge for various models of parameter setting in generative syntax, namely a scenario where all input to the learner underdetermines the target parameter setting. This scenario is exemplified by the case of zero-derived causatives in English, as discussed in Pylkkänen (2008). I then propose a model for parameter setting that uses a simple Bayesian learning procedure to learn from implicit negative evidence and arrive at the target parameter setting.

1 Introduction

An important question in language learnability is how to converge on a target grammar when all relevant grammars are compatible with the input. Indeed, this is a general challenge for various prominent models of parameter setting in generative syntax (e.g. Gibson and Wexler, 1994; Sakas and Fodor, 2001; and Yang, 2002). Consider a binary Parameter P concerning the complement of a head X^0: the complement could simply be YP (1a) or the more complex ZP containing YP (1b).

(1) a. $[_{XP}$ X [YP]]
     b. $[_{XP}$ X $[_{ZP}$ Z $[_{YP}$ Y]]]

Further, suppose that the target setting for a learner is the simpler structure in (1a), but that all the input the learner receives is ambiguous as to the parametric choice in (1). In such a case, we can ask how the learner can be sure to arrive at the adult grammar of (1a). In this paper, I present a simple case study that illustrates the learning challenge in (1) with zero-derived causatives (ZDCs) in English under Pylkkänen's (2008) theory of causatives. I propose a Bayesian model for parameter setting that learns the target setting from implicit negative evidence: given repeated instances of ambiguous input, the structure in (1a) has a greater likelihood of being the correct analysis. This result is a consequence of the learning process itself, and there is thus no need to invoke some principle such as the Subset Principle (Berwick, 1986), or to resort to default values for parameter setting.

2 The Learning Challenge with Zero-derived Causatives

Pylkkänen observes that examples like (2a) are not ambiguous: only the causer John can be characterized by gumpiness, not the causee Bill. This contrasts with (2b), in which Bill's action can be characterized by grumpiness.

(2) a. John awoke Bill in a state of grumpiness.
       ✓John is grumpy (high reading)
       ✗Bill is grumpy (low reading)
   b. Bill awoke in a state of grumpiness.

The question Pylkkänen asks is: if we follow Parsons (1990) in assuming that causatives involve a causing and caused eventuality, why do the PPs in (2a) unambiguously modify the causing event (and thus the state of the causer) and not the caused eventuality?[1] I call the possible adverbial interpretation in (2a) the high reading, and the impossible

---

[1] Thus I assume that such adverbials can modify eventualities but not nominal arguments such as the subject or object.

interpretation the low reading. Pylkkänen concludes that the lack of a low reading in (2a) is due to a structural property of the causatives. If we follow Pylkkänen, then with respect to learning we can ask how the learner learns this structural property such that there is no ambiguity in (2a).

Pylkkänen assumes there is a Cause-head in the syntax that introduces a causing event, which is phonologically null in ZDCs, and claims that there is parametric variation as to what the complement of the Cause-head is. This is the Cause-selection Parameter, which can account for cross-linguistic variation in causative structures.[2] For the sake of discussion, I will limit the range of complements to a binary choice, though the model could be expanded to accommodate the full range of parameter values Pylkkänen proposes. The choices the learner considers here are Root-selecting or Verb-selecting, schematic structures of which are in (3).

(3) a. *Root-selecting Cause*

$Cause_v$ $\sqrt{ROOT}$
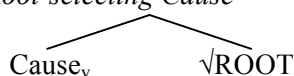
   b. *Verb-selecting Cause*

$Cause_v$

       v $\sqrt{ROOT}$
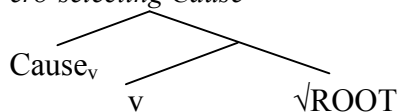
In both structures there is a category neutral lexical root that is embedded by the Cause-head. For ZDCs in English, this root could be $\sqrt{BREAK}$ or $\sqrt{MELT}$ and will be verbalized by a category-defining head. (See Borer, 2005 for discussion of category-neutral roots and category-defining morphology.) And in both structures, it is the head immediately above the root that verbalizes it. Before verbalization, though, the root combines with the internal argument and projects a $\sqrt{P}$. The difference in the Cause-selection Parameter in (3) can be thought of as a difference in which functional head verbalizes the $\sqrt{P}$. Is it simply a category-defining little v^0 with no apparent semantic contribution (which can also be phonologically null), or is it the Cause-head, which is a flavor of little v^0

itself? The difference might appear to be slight, but a Verb-selecting parameter setting crucially results in a more permissive grammar, allowing for more modification possibilities. Cross-linguistic variation with respect to modification possibilities is then the result of a language's choice in Cause-selection for a particular causative morpheme. Under both hypotheses, though, the external argument for English ZDCs would be in the specifier position of CauseP.

In light of the structures in (3), I return to the lack of ambiguity in (2a). According to Pylkkänen's argumentation, modifiers such as the PP in (2a) are verbal modifiers. That is, they can syntactically attach to verbal projections, but because they are not root modifiers, they cannot attach to the $\sqrt{P}$. With a Root-selecting causative, there is only one verbal attachment site, namely adjoining to CauseP in (3a). In contrast, a Verb-selecting causative provides two verbal attachment sites in (3b): adjunction to the vP of the verbalizing little v^0 and adjunction to the CauseP. The fact that Verb-selecting Cause provides more options for adjunction corresponds to a difference in interpretive possibilities for the two structures. When the Cause-head is merged in the derivation, the caused eventuality is existentially closed. Pylkkänen's argument is based on the following assumption about how event semantics are computed: when the lower caused eventuality is existentially closed, eventuality modifiers adjoined to CauseP can modify only the higher causing event introduced by the Cause-head. Thus lower modification of the caused eventuality by verbal modifiers is simply impossible in (3a), and this is an immediate consequence of the structure, given that there are no verbal projections below the Cause-head. In the structure for Verb-selecting Cause in (3b), though, modification of the lower caused eventuality is possible just in case the verbal modifier adjoins to the lower vP projection. The only way for the low reading in (2a) to be possible, then, involves adjunction to vP in (3b). But given that the low reading is not available in the causative in (2a), Pylkkänen concludes that there must be no vP projection in the structure of the ZDCs, a criterion that can be satisfied only with Root-selecting cause. Thus the simpler syntactic structure of Root-selecting cause in ZDCs derives the lack of ambiguity with verbal modifiers in (2a).

---

[2] Note that this parameter is relative to a particular morpheme in a language. Thus, if a language has multiple causative morphemes, each one's setting for this parameter could be different. I will discuss parameter setting only for ZDCs in English. Further, I will assume that setting this parameter for ZDCs is independent of setting any other syntactic parameters.

Turning to a learning perspective of Pylkkänen's argument, the adult grammar, which allows the high reading in (2a), can be taken to be the target state for the learner's grammar; this target state will be taken as evidence that the learner has the correct parameter setting. Pylkkänen's claim is that examples such as (2a) show that ZDCs in English are Root-selecting and thus instantiate the simpler structure in (3a). Assuming Pylkkänen is correct, the central empirical concern of this paper concerns learning a parameter setting of Root-selecting (3a) over that of Verb-selecting (3b) for these causatives in English.

The core data Pylkkänen presents for a Root-selecting setting in English ZDCs is of the sort in (2a), but the challenge for the learner is that this input underdetermines which analysis (Root or Verb-selecting) is the correct parameter setting. Consider again the example in (2a), repeated here:

(4) John awoke Bill in a state of grumpiness.

In order for a grammar to account for such an example, it must be able to generate a string-meaning pair that (among other things) (a) has a Cause-head that embeds a root and (b) has the modifier adjoin to CauseP, thereby modifying the causing event. A grammar with a parameter setting of either Root-selecting or Verb-selecting Cause is able to generate such output as is clear from the preceding discussion. Note that the same parametric ambiguity is true for the non-modified examples in (5).

(5) John awoke Bill.

To generate the example in (5), the grammar does not even need to consider which projection an adverb is adjoining to and which eventuality it is modifying – the two parameter settings are seemingly equally good at providing Cause-heads that embed lexical roots.

Recall that Pylkkänen's argument crucially involves considering the *impossibility* of the low reading (i.e. negative data),[3] a reading that a child

will presumably never be exposed to in the primary linguistic data. Given that there is no clear positive evidence in favor of the Root-selecting hypothesis, we are left with the following acquisition challenge: how do children correctly choose between Root-selection and Verb-selection for the Cause-selection Parameter? Pylkkänen's argument relies on negative evidence, but how can children learn from this evidence? I note that the learner is now faced with an instantiation of the learning challenge sketched in (1).

Before proposing a learning model that addresses this challenge, and which crucially capitalizes on the fact that a learner never hears low adverbial modification, I frame the learning challenge in the context of the 'Subset Principle' (Berwick, 1986). If we consider the structural and interpretive properties of the two causative structures in (3), we see that those of Root-selecting Cause are a proper subset of those of Verb-selecting Cause. Thus (a) the core set of syntactic heads is {Cause$_v$, √} for Root-selecting and {Cause$_v$, v, √} for Verb-selecting; (b) the set of verbal adjunction positions is {CauseP} for Root-selecting and {CauseP, vP} for Verb-selecting; and (c) the set of interpretive possibilities for verbal modifiers is {high-reading} for Root-selecting and {high-reading, low-reading} for Verb-selecting. One way to state the Subset Principle would be the following: given two hypotheses X and Y such that X can be considered a proper subset of Y, do not consider Y unless forced to do so by the input. If we consider the simpler structure of Root-selecting Cause to be a subset of the more complex structure of Verb-selecting Cause, and given that both structures adequately account for the modified and non-modified data in (4) and (5), one could invoke the Subset Principle as follows. Children learning ZDCs in English only ever consider the simpler Root-selecting structure, and are never forced to consider the more complex Verb-selecting struc-

---

[3] Pylkkänen also claims that the absence of ZDCs that have unergative counterparts with the same root is also evidence for a Root-selecting setting. This claim is based on the assumption that such ZDCs are structurally impossible given a Root-selecting setting. This is a difficult claim to evaluate. First, it is not entirely clear in Pylkkänen's analysis why such ZDCs would be ruled out structurally. Second, the absence of such

verbs is questionable. The interested reader is invited to apply the tests for unaccusativity/unergativity in Levin and Rappaport Hovav (1995) to verbs such as *graze* and *choke*. These verbs pattern as unergatives and not unaccusatives, but have ZDC forms. Nevertheless, to the extent that Pylkkänen's claim is correct, the absence of these ZDCs would constitute another form of implicit negative evidence that could be incorporated into the model. Having two kinds of implicit negative evidence (i.e. absence of low adverbial modification and of ZDCs with unergative counterparts) would presumably assist the model in the learning task.

ture (because, for example, they never hear such a causative with a low reading, which cannot be generated with the Root-selecting structure).

A similar point also holds for a default parameter setting. One could suppose that children have a default parameter setting of Root-selecting that is only switched to Verb-selecting given appropriate triggering input (such as adverbial modification of the caused eventuality).

A contribution of the learning procedure I propose is that the simpler or 'subset structure' can be learned without needing to invoke either a principle that achieves this result or a default parameter setting.
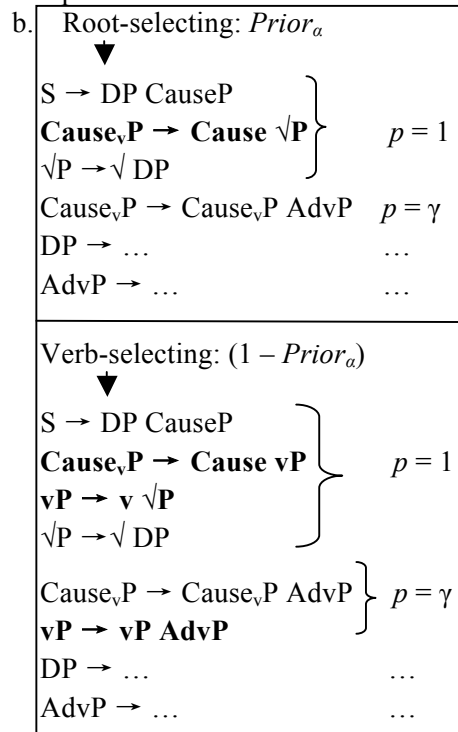
3   A Model for the Learning Challenge

The core insight of the Bayesian model proposed here is that the learner is sensitive to the *absence* of verbal modification. In the more complex Verb-selecting grammar there is a greater expectation or probability that such evidence will occur. Given that such evidence does not occur more frequently than expected under the Root-selecting grammar, the more complex grammar will leak probability, and the learning process will ultimately settle on the simpler structure, for which there is no such expectation.

I will take a learner's grammar to be a probabilistic generative model. This means the learner will take input from the primary linguistic data and try to output a string-meaning pair that matches that input as closely as possible. The way the output is generated is determined by a number of probabilistic choices. The Cause-selection Parameter can be represented as one of these choices. If these choices generate the target output, the probability distributions of these choices will be updated so as to maximize their being chosen again given similar input.

Let us consider how the model might generate input such as (5). We can base the model's learning on the rather commonplace example in (5), thereby generalizing the source of implicit negative evidence from the presumably infrequent example of the sort in (2a) that Pylkkänen discusses. I will represent the choice-points in the model as hierarchical phrase structure rules (PSRs) as in a PCFG (cf. Perfors et al., 2006). Assuming the only necessary difference between a Root and Verb-selecting grammar is the choice for the Cause-selection Pa-

rameter, this parameter can be placed on a higher tier than the other PSRs. These choice-points are all associated with priors. A schematic representation is given in (6), assuming a simplified syntax with a minimal number of PSRs. Crucially, there are PSRs for adverbial modification of CauseP and vP, which I assume are equally likely to be modified; these reflect the learner's expectation that any syntactic projection can be modified.

(6) a. Input: *John awoke Bill.*

b. Root-selecting: $Prior_\alpha$
▼

$S \rightarrow DP\ CauseP$
$Cause_vP \rightarrow Cause\ \sqrt{P}$ $\Big\}$ $p = 1$
$\sqrt{P} \rightarrow \sqrt{}\ DP$

$Cause_vP \rightarrow Cause_vP\ AdvP$ $p = \gamma$
$DP \rightarrow \ldots$ $\ldots$
$AdvP \rightarrow \ldots$ $\ldots$

Verb-selecting: $(1 - Prior_\alpha)$
▼

$S \rightarrow DP\ CauseP$
$Cause_vP \rightarrow Cause\ vP$
$vP \rightarrow v\ \sqrt{P}$ $\Big\}$ $p = 1$
$\sqrt{P} \rightarrow \sqrt{}\ DP$

$Cause_vP \rightarrow Cause_vP\ AdvP$ $\Big\}$ $p = \gamma$
$vP \rightarrow vP\ AdvP$
$DP \rightarrow \ldots$ $\ldots$
$AdvP \rightarrow \ldots$ $\ldots$

A few comments on (6) are in order. The PSRs are admittedly a simplification of English syntax – I abstract away from additional functional projections such as CP and TP (i.e. $S \rightarrow DP\ CauseP$), and do not fully expand some phrasal nodes (e.g. DP), or include terminal nodes (e.g. *Bill*) – but they allow the model to distill what is essential in the learning challenge. I thus abstract away from all PSRs between the two grammars other than choice of Cause-head and adverbial modification. By hypothesis, these other choices are identical across the two grammars, and abstracting away from them allows us to focus on learning the Cause-selection Parameter. In a sense then, these PSRs have been reverse-engineered to streamline the learning process here. Further, in the spirit of this simplicity, the corpus that the model learns from will contain

only utterances of the form in (6a). I confine myself to such a pared-down model so as to focus on the learning challenge introduced in (1), though a scaled-up model with an enriched corpus and set of PSRs should not crucially change any fundamental issues under discussion.

We can now consider the priors for the probabilistic differences between the two grammars, namely the choice of Cause-head and adverbial modification. I assume that the priors for Root- and Verb-selecting grammars are sampled from a dirichlet distribution with initial pseudo-counts of (1, 1). For the likelihood that any verbal projection is adverbially modified, γ, we could approximate it via a frequency rate of sampled verbal projections from a corpus. So long as $0 < \gamma < 1$, the actual value of γ is immaterial; it suffices to illustrate the workings of the model to simply plug in various probabilities for this value.

Before discussing the update procedure for posterior probabilities, we can now see how the more permissive Verb-selecting grammar will leak probability given the input. The probability of generating non-modified output given the Root-selecting grammar (G_{Root}) is the joint probability of choosing the Root-selecting grammar and choosing no adverbial modification at the CauseP phrase marker, as shown in (7a).

(7) a. $p(G_{Root})$ =

　　　$p(\neg CausePAdvP|CauseP)$ *
　　　$p([Cause \sqrt{P}]|CauseP)$ =

　　　$(1 - \gamma)$ * $(Prior_a)$

　b. $p(G_{Verb})$ =

　　　$p(\neg CauseP\ AdvP\ |CauseP)$ *
　　　$p([Cause\ vP]|CauseP)$ *
　　　$p(\neg vP\ AdvP|vP)$ =

　　　$(1 - \gamma)$ * $(1 - Prior_a)$ * $(1 - \gamma)$ =

　　　$(1 - \gamma)^2$ * $(1 - Prior_a)$

In contrast, the probability of generating non-modified output under the Verb-selecting grammar (G_{Verb}) is the joint probability of choosing the Verb-selecting grammar and choosing no adverbial modification at both the CauseP and vP levels (7b).

Given initial pseudo-counts of (1, 1), with repeated sampling the average probability of choosing either Cause-head will be approximately equivalent; thus the probability of not having a vP modifier causes the Verb-selecting grammar to leak probability, resulting in the probability of the data being greater under the Root-selecting grammar. This push toward Root-selecting is amplified under the update procedure with multiple tokens of input.[4]

As an update procedure, I assume that the totals for the number of times each Cause-head is sampled while successfully generating target output are used as new pseudo-count values in the dirichlet distribution. Suppose the model runs until successfully generating target output 500 times. Next, suppose that of those 500 times, Root-selecting cause was sampled 300 times, and Verb-selecting cause 200. The new pseudo-counts will then be (300, 200). These new pseudo-counts represent revised expectations about the likelihood of each grammar generating the target output.

Finally, consider how the model learns upon receiving additional input. In the case of a second input sentence, the model will now use the updated pseudo-counts from generating output conditioned by the first input token. The model will next generate 500 times the entire corpus it has been exposed to. This means that each time that the model now chooses a grammar (based on repeated sampling of the updated dirichlet distribution), it will try to use that grammar and all subsequent choices dependent on that grammar to generate both the original first token of input and the second token as output.

Thus when the model encounters $n > 1$ tokens of input, the model will (a) take the sums of successes per grammar with $(n - 1)$ tokens of input and use these sums to update the pseudo-counts of the dirichlet distribution; then (b) generate the entire corpus of n tokens of input 500 times using posterior probabilities from the updated dirichlet distribution. This process repeats until only a single parameter setting is used to generate the entire corpus, at which point the model can be said to have learned that parameter setting. In this way, the model benefits from rapid and efficient learning from a small amount of input data. This rapid learning has been illustrated in numerous cognitive

---

[4] Note that although (7) has the effect of making the subset grammar more likely to generate target output, it is not another version of the Subset Principle. Rather (7) reflects the more general mechanisms of how a PCFG can generate output.
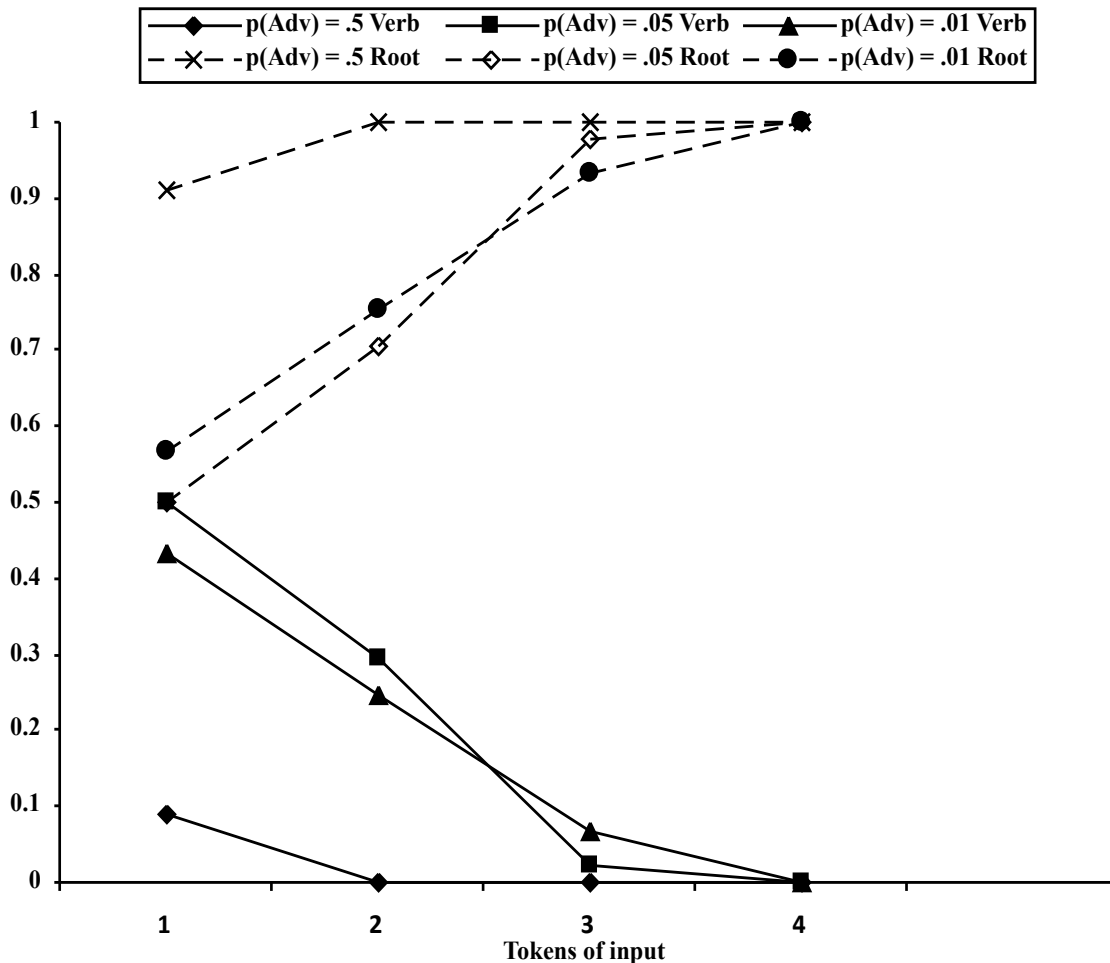
Figure 1. Average success-rate per grammar for target output

experiments outside the domain of language and has been modeled in a Bayesian framework (Kemp et al., 2007).

Indeed, sample results from running the model indicate its success at learning the target Root-selecting setting given a small input corpus. Simulations of the model were run with a simple program written in the Church language (Goodman et al., 2008). The results reported here are the average probability for each grammar being chosen given the output matching the attested input after running the model 10 times. The results are given in Figure 1 in a time-course graph showing averages for different amounts of input data, which reflect the effect of updating the priors. As the probability of a verbal projection being modified has been left as a variable, Figure 1 shows various representative values. Each graph-line shows the average success-rate of a certain grammar given a particular prob-

ability for adverbial modification of a verbal projection under that grammar. For example, *p(Adv) = .5 Verb* corresponds to a line representing the average percentage of the time the Verb-selecting setting was chosen from among the target output, given that the probability of verbal modification was .5.

What Figure 1 shows is that after only a few tokens of input, the Root-selecting grammar is overwhelmingly the more likely option. If the probability of verbal modification is .5, then the success-rate of the Root-selecting grammar is 1 after 3 tokens of input, while that of the Verb-selecting grammar is 0. This is surely an unrealistic probability to have for verbal modification, but even if we decrease it to .05 or .01 the model still settles on the Root-selecting grammar. With a smaller probability for verbal modification, it now takes the model 4 tokens of input before the suc-

cess-rate of the Verb-selecting grammar reaches or approaches 0. In fact, the best that the Verb-selecting grammar does is an average success-rate of .0008 (.9992 success-rate for Root-selecting) when the probability of verbal modification is .01.

These results clearly show that the model is learning the Root-selecting grammar as the correct parameter to generate target output. Further, the model is able to learn on the basis of as few as 4 tokens of input. Going beyond the baseline model presented here, to the extent that the priors are on the right track and that the probability of verbal modification is reflective of expanded corpus results, the prediction is that expanded versions of the model will also be successful.

4 Comparison with Other Models

In this section I briefly compare the Bayesian model proposed here with three prominent models that attempt to learn correct syntactic parameter settings: Yang (2002), Gibson and Wexler (1994), and Sakas and Fodor (2001). None of these three models can guarantee convergence on the target Root-selecting setting for ZDCs. For the sake of comparison, keeping to a corpus like (6a), let us assume that in all models we have a binary parameter such as Root- or Verb-selecting cause, and that the choice of this parameter has no effect on any other parameter setting.

The core of Yang's (2002) probabilistic learning model involves increasing or decreasing a parameter's probability based on whether adopting that parameter leads to a grammar that is compatible with the input data. Thus whenever the model encounters any data containing ZDCs, it will sample a Cause-head parameter setting based on the probability distribution and test out this setting to see whether it is compatible with the input. Yang explicitly discusses how his model is not reliant on what have been called unambiguous triggers in Fodor (1998). An unambiguous trigger would be a token of input data that is compatible with only a single (relevant) parameter setting, thereby excluding all other relevant parameter settings. In the discussion on causatives above, an unambiguous trigger would be input that showed the availability of the low adverbial reading: this input is compatible only with the Verb-selecting hypothesis and not with the Root-selecting hypothesis. However, implicit in Yang's discussion is that for each non-

target parameter setting there must be some input that is not compatible with it. As long as such input exists, it will result in the non-target parameters being punished, and so long as these non-target parameters are punished sufficiently, in the long run the target parameter setting will eventually prevail.

The scenario of ZDCs in English, then, is problematic for Yang's model. All the relevant parameter settings are compatible with the input, and there is thus no input data that can rule out any of the parameter settings. As Root and Verb-selecting parameter settings will have similar reward-punishment rates in this situation, all things being equal (e.g. non-biased priors), the model could converge on either setting or get stuck in a state of stasis, with neither setting's probability exhibiting asymptotic behavior (cf. discussion in Pearl, 2009). Compared to the model proposed in this paper, Yang's model is unable to learn from implicit negative evidence: unlike the Bayesian model, Yang's model does not go beyond grammar compatibility to consider the probability of the data given a particular grammar.

Similarly, in the error-driven model of Gibson and Wexler (1994), there is no guarantee that the learner will converge on the target parameter setting for ZDCs. In this model, parameter settings have weights of 1 or 0, and a parameter's value is changed only if the current vector of parameters is incompatible with the most recent token of input. In such a case, only one parameter can be changed (the Single Value Constraint). Which parameter is chosen to have its value changed is left as an open question, but there is a constraint such that whatever the new parameter vector is, the grammar represented by that new vector must now be compatible with the most recent input (the Greediness Constraint).

Consider, then, how the Gibson and Wexler model fares if the initial state, which is some random grammar or parameter vector, has a non-target parameter setting for English ZDCs. No input containing a ZDC could force the Cause-selection Parameter to change its value because both settings are compatible with that data. Further, even if this input forced the model to change its current grammar (because of non-target setting of some other parameter), the model would not change the setting of the Cause-selection Parameter because no new value for this parameter would help in the face of

the latest input (the Greediness Constraint). The model would have to change the value of some other parameter and leave the Cause-selection Parameter alone (the Single Value Constraint). Thus the model will be in a local maximum: no input could push the model toward a target setting for ZDCs, and the model would remain stuck in a non-target setting. Of course, if the initial state was a Root-selecting grammar, then no input in English would push the learner from that setting, and the learner would have the target parameter setting.

Finally, the model in Sakas and Fodor (2001) crucially relies on input that contains the unambiguous triggers discussed above. In their model, as the parser builds a parse tree of the input, the parser is able to recognize at any point in the structure whether a parametric choice is underdetermined given the input data. For the case of the ZDCs discussed in this paper, the parser, upon facing the Cause-head in the parse tree, presumably would be able to determine that either a vP or √P complement is compatible with the input data. In the terms of Sakas and Fodor, the parser is faced with an ambiguity with respect to parameter selection. What the parser then does is report this ambiguity to the learning mechanism. The learning mechanism will then not use this 'ambiguous input' to learn a parameter setting. In other words, the learning mechanism will wait until an unambiguous trigger occurs in the input before setting any parameter value. Now as we have discussed, all the relevant data for zero-derived causatives in English underdetermine the correct structural analysis – it is all ambiguous input, and there is no unambiguous input. As it stands then, Sakas and Fodor's model is unable to learn the correct parameter setting when faced with the challenge of ZDCs.

Before closing this section, I note that an amendment to both Gibson and Wexler's and Sakas and Fodor's models would be able to account for the Cause-selection Parameter: a default parameter setting. The learning mechanism would only need to consider other parameter settings if pushed toward them by the input. If Root-selecting Cause was the default value, then the English zero-derived causatives would be accounted for. Only if the input data presented some evidence that is incompatible with a Root-selecting parameter setting (e.g. an utterance with the low adverbial reading) would the learning mechanism change from the default to a Verb-selecting setting. As mentioned

in the introduction, though, an advantage of the model here is that no default needs to be specified.

5 Concluding remarks.

I have introduced a Bayesian model that is up to the learning challenge that Pylkkänen's theory of parameters presents us with for the case of English ZDCs in English. Given input that underdetermines that correct structural analysis, the model is able to learn from implicit negative evidence with respect to the likelihood of verbal modification and select the correct, simpler, and more restrictive parameter setting. No default value for the parameter setting was necessary, nor any principle such as the Subset Principle. The model is a simple illustration of the how the learning procedure itself in a Bayesian framework results in the correct parameter setting. Further, other prominent models of parameter setting are not capable of learning the correct parameter setting given the underdetermining nature of the data. To be sure, the model is only the simplest illustration of how this learning procedure works, and a clear direction of future research can focus on expanding its empirical scope. Now that the model has success at the most basic level we can consider scaling it up. One way to expand is to enlarge the corpus that is used as input data so that it better approximates input that a child encounters.[5] Another consideration concerns learning a Verb-selecting grammar in languages where the low reading is possible. In the absence of input with adverbials in the corpus, the model here predicts that only the Root-selecting grammar will be learned. This suggests there must another property in the input to allow for learning a Verb-selecting grammar in languages that have it; this could be a morphologically overt v^0 between CauseP and √P. Indeed, all the Verb-selecting languages discussed in Pylkkänen have such overt morphology. Such intervening morphology is impossible in Root-selecting languages, and true to their name, ZDCs in English display no such head.

---

[5] This could include input tokens with verbal modification, a very high proportion of which could push the learner toward the more complex Verb-selecting grammar. This is because the probability of modifying CauseP *or* vP given Verb-selecting is greater than that of just modifying CauseP given Root-selecting. Given a high enough proportion of the input containing verbal modifiers, this could swing the balance of data in favor of a Verb-selecting setting. It is doubtful, though, whether learner input actually contains such a high proportion.

The non-deterministic nature of the model also means there is a developmental implication for language acquisition in children: at earlier stages in the learning procedure, non-target parameter settings with likelihoods that are not too low are viable choices. Before parameter setting is finalized, then, we might expect non-target behavior from children with respect to, say, the Verb-selecting parameter setting (see Yang 2002 for discussion of this point). Is there evidence that children sometimes treat zero-derived causatives in English as being Verb-selecting before having learned that they are in fact Root-selecting? The model would lead us to expect that in initial stages of learning, the likelihood of a Verb-selecting analysis is high enough that children would incorrectly treat them as being Verb-selecting at least some of the time. Careful experimental work would be needed to test these predictions, but to the extent that they are borne out, in addition to showing how target parameter settings can be learned, an advantage of the non-deterministic framework here is its potential to model non-target behavior.

Finally, a contribution of this paper is to add to the emerging body of literature incorporating Bayesian modeling into generative linguistics. As illustrated in Pearl and Goldwater (in press), though, much of this has not looked at setting syntactic parameters. A notable exception is the line of research initiated by Regier and Gahl (2004), which attempts to learn the syntactic structure and semantics of anaphoric *one* in English. The learning issues related to anaphoric *one* differ from those of ZDCs here in at least two important ways. As Payne et al. (2013) note, (a) not all input the learner receives concerning anaphoric *one* is ambiguous, and (b) the properties that the model attempts to learn reflect only preferences in the adult grammar. The case of ZDCs, then, presents a learning model with an ideal test of the learning challenge presented in (1): categorical parameter setting in the face of entirely ambiguous evidence.

Acknowledgments

References

Berwick, Robert C. 1986. Learning from positive-only examples: The subset principle and three case studies. In R. S. Michalski, J. C. Carbonell, and T. M. Mitchell (eds.), *Machine learning: An artificial intelligence approach, Vol 2*. Los Altos: Morgan Kaufmann.

Borer, Hagit. 2005. *Structuring Sense, Vols. 1-2*. Oxford: Oxford university Press.

Fodor, Janet Dean.1998. Unambiguous Triggers *Linguistic Inquiry* 29 (1): 1-36.

Gibson, Edward and Ken Wexler. 1994. Triggers. *Linguistic Inquiry* 25 (3): 407-454.

Goodman, Noah, Vikash Mansinghka, Daiel Roy, Keith Bonawitz, and Joshua Tenenbaum. 2008. Church: A language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Kemp, Charles, Amy Perfors, and Joshua Tenenbaum. 2007. Learning overhypotheses with hierarchical Bayesian models. *Developmental Science* 10 (3): 307-321.

Levin, Beth and Malka Rappaport Hovav. 1995. *Unaccusativity*. Cambridge: MIT Press.

Parsons, Terence. 1990. *Events in the semantics of English: A study in subatomic semantics*. Cambridge: MIT Press.

Payne, John, Geoffrey Pullum, Barabara Scholz, and Eva Berlage. 2013. Anaphoric *one* and its implications. Language 89 (4): 794-829.

Pearl, Lisa. 2009. *Necessary bias in natural language learning*. PhD dissertation, The University of Maryland.

Pearl, Lisa and Sharon Goldwater. In press. Statistical learning, inductive bias, an Bayesian inference in language acquisition. In J. Lidz, W. Snyder, and C. Pater (eds.), *The Oxford Handbook of Developmental Linguistics*.

Perfors, Amy, Joshua Tenenbaum, and Terry Regier. 2006. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Pylkkänen, Liina. 2008. *Introducing Arguments*. Cambridge: MIT Press.

Regier, Terry and Susanne Gahl. 2004. Learning the unlearnable: The role of missing evidence. *Cognition* 93: 147-155.

Sakas, William G. and Janet Dean Fodor. 2001. The Structural Triggers Learner. In Stefano Bartolo (ed.), *Language acquisition and learnability*, 172-233. Cambridge: Cambridge University Press.

Yang, Charles. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.