Modeling the learning of the Person Case Constraint

Adam Liter
Dept. of Linguistics
University of Maryland
College Park, MD 20742 USA
io@adamliter.org

Naomi H. Feldman
Dept. of Linguistics & UMIACS
University of Maryland
College Park, MD 20742 USA
nhf@umd.edu

Abstract

Many domains of linguistic research posit feature bundles as an explanation for various phenomena. Such hypotheses are often evaluated on their simplicity (or parsimony). We take a complementary approach. Specifically, we evaluate different hypotheses about the representation of person features in syntax on the basis of their implications for learning the Person Case Constraint (PCC). The PCC refers to a phenomenon where certain combinations of clitics (pronominal bound morphemes) are disallowed with ditransitive verbs. We compare a simple theory of the PCC, where person features are represented as atomic units, to a feature-based theory of the PCC, where person features are represented as feature bundles. We use Bayesian modeling to compare these theories, using data based on realistic proportions of clitic combinations from child-directed speech. We find that both theories can learn the target grammar given enough data, but that the feature-based theory requires significantly less data, suggesting that developmental trajectories could provide insight into syntactic representations in this domain.

1 Introduction

Representing surface realizations as bundles of features is ubiquitous in linguistics. For example, in syntax, different forms that result from subject-verb agreement are taken to be the result of different feature bundles. Relevant features for subject verb agreement in English include at least the tense and the number of the subject. Although there is little variation in the different surface forms for English verbs, the verb *walk* does differ when the subject is singular and the tense is present (*walks*), compared to when the subject is singular and the tense is past (*walked*).

Features are often taken to be either privative or binary (though these are not the only possibilities). For example, some might argue that the English singular/plural distinction is based on a privative feature: a noun phrase can either be specified as plural or not specified for number (*e.g.*, [plural] and [ ]). In this case, when *dog* is marked with "[plural]", it is realized as *dogs*. Others might argue that the distinction is based on a binary feature: a noun phrase can be specified as "plus" or "minus" (*e.g.*, [+plural] and [−plural]). In this case, when *dog* is marked with "[+plural]", it is realized as *dogs*.

Feature representations are typically evaluated based on the extent to which they simplify linguistic analyses, that is, on their ability to provide parsimonious descriptions of cross-linguistic grammatical patterns. For a concrete example of this type of argument, see Adger and Smith (2010), who argue that both the intra-dialectal variation in the inflection of the verb *be* in Buckie Scottish English as well as the inter-dialectcal variation in the inflection of the verb *be* in English more broadly is nicely explained by a feature system involving binary-valued features of Singular, Participant, and Author.

In this paper, we take a different approach to evaluating feature representations, focusing on their implications for learning (for similar approaches, see Pearl and Sprouse, 2013; Pearl et al., 2017; Rasin and Katzir, 2017; Pearl and Sprouse, 2019). Specifically, we investigate how person features might be represented in the syntactic component of the grammar, using the domain of clitics as a case study and the learnability of a phenomenon involving clitics as a metric for plausibility. We find that both of the representational theories that we test can learn the target grammar given enough data, but that they differ considerably in the amount of data they require. This suggests that children's learning trajectory has the potential to provide insight into syntactic representa-

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 372-381.
New Orleans, Louisiana, January 2-5, 2020

tions in this domain.

2 The Person Case Constraint

Clitics are bound morphemes (*i.e.*, morphemes that cannot stand on their own). The clitics relevant to the PCC are pronominal clitics, which encode first, second, and third person, and they must occur immediately next to a free morpheme, usually a verb. For example, (1) shows a Spanish sentence where the direct and indirect objects of a ditransitive verb are both realized as clitics. The clitics are immediately before the verb, and, in this case, they encode first and third person, respectively.

(1)	Me	lo	cuentas
	1.SG.DAT	3.SG.ACC	tell
	‘(You) tell it to me’		

Interestingly, when the direct and indirect objects to a ditransitive verb are both realized as clitics, not all combinations are possible. For example, compare (1) to (2), where the first person clitic serves as the direct object and the third person clitic serves as the indirect object (*i.e.*, the opposite of (1)). The sentence in (2) is ungrammatical.[1]

(2)	* Me	le	recommendó
	1.SG.ACC	3.SG.DAT	recommend.PST
	‘S/he recommended me to her/him’		

The ungrammaticality of (2) is part of a broader phenomenon called the Person Case Constraint (PCC) (see, *e.g.*, Bonet, 1991, 1994); the PCC will be the central focus of our case study on the representation of person features.

Ignoring the possible combinations of direct and indirect objects with either both first person or both second person arguments[2] gives seven different possible direct and indirect object pairings: 1 2, 1 3, 2 1, 2 3, 3 1, 3 2, and 3 3. There are four attested variants of the PCC, each of them banning a different subset of these seven possible clitic combinations. The four variants of the PCC (and their names) are given in Table 1, along with languages/dialects that are known to instantiate each of them (note that these tables include 3 3 and thus differ slightly from those reported in Graf, 2012, p. 86).

Because there are different variants of the PCC that occur cross-linguistically, a child will have to learn which variant their language instantiates on the basis of input.

3 Evaluating two theories of the PCC

We use a Bayesian learning model to evaluate the plausibility of two theories of the representation of person features. The first theory is one in which first, second, and third person have no further structure; they are just represented as atomic features in the grammar, like in (3). We refer to this as the simple theory of the PCC because the grammar is assumed to simply state, for each possible clitic combination, whether it is grammatical.

(3) a. $1 = 1$
b. $2 = 2$
c. $3 = 3$

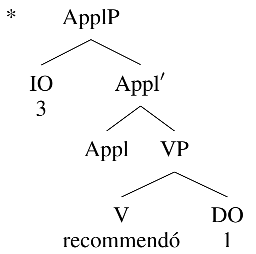
We compare this to another theory in which first, second, and third person are represented as feature bundles, consisting of two values, one for the binary feature Author and one for the binary feature Participant, as in (4) (Nevins, 2007). We refer to this as the feature-based theory of the PCC.

(4) a. $1 = \begin{bmatrix} +\text{Auth} \\ +\text{Part} \end{bmatrix}$

b. $2 = \begin{bmatrix} -\text{Auth} \\ +\text{Part} \end{bmatrix}$

[1] Note that even though the first person clitic occurs before the third person clitic in both (1) and (2), the literature usually talks about the ungrammaticality of (2) with the starred string “*3 1”. This is meant to indicate the underlying argument structure relations—namely, in the syntactic analysis of (2), but not (1), the dative third person argument is structurally higher than first person argument, as shown in (i). This is generally written as “*3 1”, meant to reflect the fact that the third person argument structurally precedes the first person argument, even though the surface string order of the clitics is the opposite.

(i)



Nonetheless, as (1) shows, there are some instances where the surface string order of the clitics does match the underlying argument structure relations. This depends on a variety of language specific factors, including at least the nature of the particular verb and ordering effects between some of the clitics in some languages.

[2] The combinations with both first or both second person arguments are often ignored in this literature because of other complicating factors. Specifically, these combinations are also governed by another part of the grammar, Binding Theory (see, *e.g.*, Chomsky, 1981).

c. $3 = \begin{bmatrix} -\text{Auth} \\ -\text{Part} \end{bmatrix}$

Based on corpus data from child-directed speech, we model the learning of one PCC variant in order to investigate the plausibility of these different representations of person features. The remainder of this section lays out these two representational theories in more detail.

IO↓/DO→	1	2	3
1	NA	*	✓
2	*	NA	✓
3	*	*	✓

(a) Strong PCC (Greek, Spanish, *etc.*)

IO↓/DO→	1	2	3
1	NA	✓	✓
2	*	NA	✓
3	*	*	✓

(b) Ultrastrong PCC (Classical Arabic, Spanish, *etc.*)

IO↓/DO→	1	2	3
1	NA	✓	✓
2	✓	NA	✓
3	*	*	✓

(c) Weak PCC (French, Catalan, Spanish, *etc.*)

IO↓/DO→	1	2	3
1	NA	✓	✓
2	*	NA	✓
3	*	✓	✓

(d) Me-First PCC (Romanian, Spanish, *etc.*)

Table 1: PCC varieties (rows indicate the indirect object, and columns indicate the direct object; '✓' indicates grammatical, and '*' indicates ungrammatical; for example, *1 2 is ungrammatical in Strong PCC languages but grammatical in all other PCC varieties)

3.1 A simple theory of the PCC

The simple theory of the PCC states, for each clitic combination, whether or not it is grammatical. For this theory, person features are atomic (cf. (3)), and the grammar simply states that some combinations (*e.g.*, *2 1) are banned. Given that there are 7 clitic combinations, this leads to $2^7 = 128$ possible grammars, some of which are shown in Table 2.[3]

Grammar	1 2	1 3	2 1	2 3	3 1	3 2	3 3
SG_1	✓	✓	✓	✓	✓	✓	✓
SG_2	✓	✓	✓	✓	✓	✓	*
SG_3	✓	✓	✓	✓	✓	*	✓
SG_4	✓	✓	✓	✓	✓	*	*
SG_5	✓	✓	✓	✓	*	✓	✓
SG_6	✓	✓	✓	✓	*	✓	*
SG_7	✓	✓	✓	✓	*	*	✓
SG_8	✓	✓	✓	✓	*	*	*
...	...	...	...	...	...	...	...
SG_{21}	✓	✓	*	✓	*	✓	✓
SG_{22}	✓	✓	*	✓	*	✓	*
SG_{23}	✓	✓	*	✓	*	*	✓
...	...	...	...	...	...	...	...
SG_{55}	*	✓	*	✓	*	*	✓
...	...	...	...	...	...	...	...
SG_{85}	*	✓	*	✓	*	✓	✓
SG_{86}	*	✓	*	✓	*	✓	*
SG_{87}	*	✓	*	✓	*	*	✓
...	...	...	...	...	...	...	...
SG_{128}	*	*	*	*	*	*	*

Table 2: Some of the 128 possible simple grammars (SG) for the PCC

3.2 A feature-based theory of the PCC

Nevins (2007) proposes a feature-based theory of the four PCC varieties. This theory is much more

[3] The simple grammar for the Strong PCC would be SG_{55}, the simple grammar for the Ultrastrong PCC would be SG_{23}, the simple grammar for the Weak PCC would be SG_7, and the simple grammar for the Me-First PCC would be SG_{21}.

restrictive in that it allows many fewer possible types of grammars. For this theory, it is crucial that first, second, and third person are represented as feature bundles, consisting of two binary feature values, as shown above in (4).

The features Author and Participant are taken to be primitive features in the theory of morphosyntax, and each can be valued as either $+$ or $-$.[4] Broadly, this theory relies on how these features bundles can (or cannot) co-occur with one another in concert with a syntactic operation called Agree.

To spell out the details more carefully, clitics are understood to be the morphophonological realization of a syntactic operation called Agree (see, *e.g.*, Borer, 1984). The possible grammars in this feature-based theory thus consist of different possible specifications for the feature(s) that trigger(s) Agree. Specifically, there is a syntactic probe, *v*, that, when introduced into the derivation, triggers Agree. Nevins assumes that the probe can be specified to search for either marked and/or contrastive Author and Participant features (cf. Calabrese, 1995; Nevins, 2007, p. 285–290).

The marked version of each feature is its $+$ value. A contrastive instance of the Participant feature is one that occurs in the presence of $-$Auth; when Participant occurs with +Auth, it is not contrastive because there is no possible feature bundle $\begin{bmatrix} \text{+Auth} \\ -\text{Part} \end{bmatrix}$ (cf. fn. 4). In other words, if the feature bundle contains +Auth, it must necessarily also contain +Part. A contrastive instance of the Author feature is one that occurs in the presence of +Part; *i.e.*, when you have a feature bundle that contains $-$Part, then it must necessarily also contain $-$Auth.

Given this theory of clitics and the PCC, there are then nine possible feature-based grammars (FG), which are all given the first column of Table 3. In the grammar specifications in this table, ‘u’ indicates that the probe, *v*, is looking for a feature of the type that follows the ‘u’ to Agree with.[5] Furthermore, we indicate, for example, contrastive Author as ‘uAuth/[+Part]’, which can be read as “the probe is looking for an Author feature that occurs in the context of +Part”.

Here, we walk through two example derivations. For further discussion and derivations, see Nevins (2007, p. 290–301).

Let’s first consider the clitic order *1 2, which is disallowed in Strong PCC languages. The feature specification that is claimed to give rise to Strong PCC languages is FG$_6$.

Nevins argues that there are two conditions that govern the application of Agree (2007, p. 295), Contiguous Agree and Matched Values.

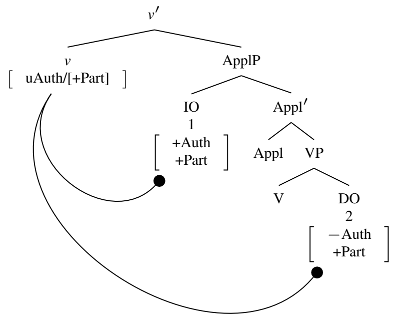
(5) Contiguous Agree: For a relativization R of a feature F on a Probe P, and $x \in$ Domain(R(F)), $\neg\exists y$, such that $y > x$ and $p > y$ and $y \in$ Domain(R(F))
“There can be no interveners between P and x that are not in the domain of relativization that includes x”

(6) Matched Values: For a relativization R of a feature F, $\exists\alpha$, $\alpha \in \{+,-\}$, $\forall x$, $x \in$ Domain(R(F)), val(x,F)$= \alpha$
“All elements within the domain of relativization must contain the same value”

In other words, Contiguous Agree requires that any argument that occurs in between the probe and the target of Agree must also itself be a target of Agree, and Matched Values requires that all arguments that are in the domain of the Agree operation must share the same value (*e.g.*, both must be +Auth; one cannot be $-$Auth and the other +Auth).

Now, in the case of *1 2 when the grammar is FG$_6$ (*i.e.*, the Strong PCC), where the probe, *v*, seeks to Agree with arguments bearing contrastive Author, a partial derivation will look like the one in (7).

(7) *



In this case, the condition Matched Values is violated. Both the first person indirect object and the second person direct object are in the domain of

[4] The feature combination of $\begin{bmatrix} \text{+Auth} \\ -\text{Part} \end{bmatrix}$ is taken to be impossible because of what the features mean—namely, it’s not possible to be the author (*i.e.*, speaker) in a conversation but not a participant in that same conversation.

[5] This is generally understood to mean “uninterpretable” in the syntactic literature; for an overview of feature theory in Minimalist theories of syntax, see Pesetsky and Torrego (2007).

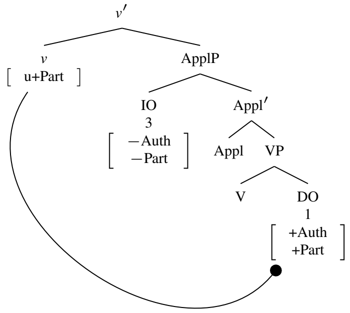
Probe	Grammar	1 2	1 3	2 1	2 3	3 1	3 2	3 3
$v\begin{bmatrix}\quad\end{bmatrix}$	FG_1	✓	✓	✓	✓	✓	✓	✓
$v\begin{bmatrix}\text{u+Part}\end{bmatrix}$	FG_2	✓	✓	✓	✓	*	*	✓
$v\begin{bmatrix}\text{u+Auth}\end{bmatrix}$	FG_3	✓	✓	*	✓	*	✓	✓
$v\begin{bmatrix}\text{u+Part}\\\text{u+Auth}\end{bmatrix}$	FG_4	✓	✓	*	✓	*	*	✓
$v\begin{bmatrix}\text{uAuth/[+Part]}\\\text{uPart/[}-\text{Auth]}\end{bmatrix}$	FG_5	*	*	*	*	*	*	✓
$v\begin{bmatrix}\text{uAuth/[+Part]}\end{bmatrix}$	FG_6	*	✓	*	✓	*	*	✓
$v\begin{bmatrix}\text{uAuth/[+Part]}\\\text{u+Part}\end{bmatrix}$	FG_7	*	✓	*	✓	*	*	✓
$v\begin{bmatrix}\text{uPart/[}-\text{Auth]}\end{bmatrix}$	FG_8	*	*	✓	*	✓	*	✓
$v\begin{bmatrix}\text{uPart/[}-\text{Auth]}\\\text{u+Auth}\end{bmatrix}$	FG_9	*	*	*	*	*	*	✓

Table 3: The 9 possible feature-based (FG) grammars for the PCC, according to Nevins (2007)

Agree for the feature uAuth/[+Part] on the probe, *v* (because they both have Author features that occur in the context of +Part). However, they have differing values for Author, so Matched Values is violated, giving rise to the ungrammaticality of *1 2 when the grammar is FG_6.

Next, let's consider the case of *3 1, which is disallowed in Weak PCC languages. The feature specification that is claimed to give rise to Weak PCC languages is FG_2, where the probe, *v*, seeks to Agree with arguments bearing a marked Participant feature. In the case of *3 1 when the grammar is FG_2, a partial derivation will look like the one in (8).

(8) *



Here, the probe is looking for a +Part feature; this means that it can agree with the direct object; however, there is a structurally higher element—namely, the third person indirect object, $\begin{bmatrix}-\text{Auth}\\-\text{Part}\end{bmatrix}$—that intervenes between the probe, *v*, and the target of Agree but is not in the domain of the probe because it does not contain a +Part feature. This violates the condition Contiguous Agree, so the clitic order *3 1 is thereby disallowed in FG_2.

Walking through the derivations for all seven possible clitic orders for all nine feature-based grammars gives the results shown in Table 3.[6]

4 The learning model

We use Bayesian modeling to implement a computational-level learning model that infers a grammar, given a bunch of sentences with ditransitive verbs and two clitics. In the case of the feature-based theory of the PCC, there are 9 grammars, and so the hypothesis space is much smaller. In the case of the simple theory of the PCC, there are 128 grammars, and so the hypothesis space is much larger.

Using realistic proportions of the occurrences of these types of constructions in child-directed speech, we seek to establish how much data would be needed to learn the correct grammar under each of these theories.

[6] The feature-based grammar for the Strong PCC would be FG_6, as noted, or FG_7 (these two feature-based grammars are extensionally equivalent), the feature-based grammar for the Ultrastrong PCC would be FG_4, the feature-based grammar for the Weak PCC would be FG_2, and the feature-based grammar for the Me-First PCC would be FG_3. The remaining grammars would then delimit the predicted typology of PCC languages. FG_1 would be a language without PCC effects (and perhaps also without clitics), like English; there would be two further predicted types of PCC languages, FG_8, which Nevins calls a "Me-Last" language, and FG_5 and FG_9, which are extensionally equivalent in only allowing 3 3 (note that Nevins (2007) does not consider 3 3 constructions).

4.1 The generative model

We assume the generative model depicted in Figure 1. A generative model encodes the assumptions a learner would have about how the data it observes are generated.

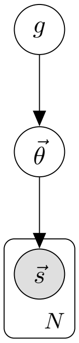


Figure 1: Generative model

Our generative model assumes that there is a grammar, g, that determines how often certain clitic combinations will be used. In the case of the simple theory of the PCC, g will be one of $SG_1, \ldots, SG_{128}$, and in the case of the feature-based theory of the PCC, g will be one of $FG_1, \ldots, FG_9$.

This grammar g is assumed to generate a vector of probabilities, $\vec{\theta}$, which governs the frequency of use of each of the different clitic combinations in the language. In other words, $\vec{\theta}$ determines how often one would expect to see each clitic combination in a corpus containing N ditransitive sentences that have cliticized both internal arguments. In our model, we assume that the elements of $\vec{\theta}$ corresponding to any clitic orderings that are disallowed under g are set to zero, and that the remaining elements of $\vec{\theta}$ are generated from a Dirichlet distribution with dimensionality equal to the number of permitted clitic orderings,

$$\vec{\theta} \mid g \sim \text{Dir}(\langle 1, \ldots, 1 \rangle) \qquad (1)$$

This Dirichlet distribution encodes a belief that any value of $\vec{\theta}$ that is consistent with the grammar is equally likely, a priori.

The instances of clitic combinations that a learner observes, represented in our generative model as $\vec{s}$, are then assumed to be sampled from $\vec{\theta}$. For example, if, in the corpus, there were 3 instances of the 1 3 clitic combination, 6 instances of the 3 3 clitic combination, and no others, then $\vec{s}$ would be $\langle 0, 3, 0, 0, 0, 0, 6 \rangle$. The generative model assumes that $\vec{s}$ are sampled from a multinomial distribution with parameter $\vec{\theta}$,

$$\vec{s} \mid \vec{\theta} \sim \text{Multinom}(N, \vec{\theta}) \qquad (2)$$

The learner observes the clitic combinations in its corpus and infers which of the possible grammars was most likely to have generated these data.

4.2 Inferring the grammar

Given a count of the occurrence of each of the seven possible clitic orders, $\vec{s}$, from a corpus of sentences, the posterior probability of each possible grammar, $p(g \mid \vec{s})$, can be computed. Using Bayes' rule, $p(g \mid \vec{s})$ can be calculated as

$$p(g \mid \vec{s}) = \frac{p(\vec{s} \mid g) p(g)}{\sum_{g'} p(\vec{s} \mid g') p(g')} \qquad (3)$$

We assume a uniform prior probability distribution over grammars, $p(g)$. The likelihood term, $p(\vec{s} \mid g)$, is calculated by integrating over all possible values of $\vec{\theta}$,

$$p(\vec{s} \mid g) = \int p(\vec{s} \mid \vec{\theta}) p(\vec{\theta} \mid g) \mathrm{d}\vec{\theta} \qquad (4)$$

Note that the complexity of each hypothesized grammar differs because in grammars that rule out some clitic combinations, the corresponding values of $\vec{\theta}$ are set to zero, and the corresponding likelihood terms have fewer values of θ to integrate over. Because of this, a grammar that allows fewer clitic combinations will have a higher likelihood than a grammar that allows more clitic combinations, when some counts in $\vec{s}$ are zero (cf. Tenenbaum and Griffiths, 2001). This is so because a more complex grammar needs to integrate over values of $\vec{\theta}$ that give probability to things that do not occur in the learner's input.

For example, in trying to determine how likely it is that g is either SG$_1$ or FG$_1$, both which allow all 7 possible clitic combinations, $p(\vec{s} \mid \vec{\theta})$ is $\frac{N!}{n_1! \cdots n_7!} \prod_{i=1}^{7} \theta_i^{n_i}$, and $p(\vec{\theta} \mid g)$ is $\frac{\Gamma(\sum_{i=1}^{7} \alpha_i)}{\prod_{i=1}^{7} \Gamma(\alpha_i)} \prod_{i=1}^{7} \theta_i^{\alpha_i - 1}$. On the other hand, if trying to determine how likely it is that g is either FG_3 or SG_{21}, both which allow 5 of the 7 possible clitic combinations, $p(\vec{s} \mid \vec{\theta})$ will be $\frac{N!}{n_1! \cdots n_5!} \prod_{i=1}^{5} \theta_i^{n_i}$, and $p(\vec{\theta} \mid g)$ will be $\frac{\Gamma(\sum_{i=1}^{5} \alpha_i)}{\prod_{i=1}^{5} \Gamma(\alpha_i)} \prod_{i=1}^{5} \theta_i^{\alpha_i - 1}$.

To calculate the likelihood that g is, for example, FG$_1$, we can substitute these terms into Eq. 4, which yields Eq. 5 (cf. Gelman et al., 2014).

$$\frac{\prod_{i=1}^{7}\Gamma(n_i+\alpha_i)}{\Gamma\left(\sum_{i=1}^{7} n_i+\alpha_i\right)}\frac{N!}{n_1!\cdots n_7!}\frac{\Gamma\left(\sum_{i=1}^{7}\alpha_i\right)}{\prod_{i=1}^{7}\Gamma(\alpha_i)} \quad (5)$$

On the other hand, if calculating the likelihood that g is instead FG_3, then all of the instances of '7' in Eq. 5 would be replaced with '5'.

Having defined the learning model, we can now give it data to learn from, based on child-directed speech, and see what difference the size of the hypothesis space makes.

5 Simulations

We conducted several simulations based on realistic proportions of clitic combinations taken from child-directed speech.

5.1 Data

We estimated the frequency of each clitic combination in child-directed speech based on their distribution in the Aguirre Corpus (Aguirre, 2003), from CHILDES (MacWhinney, 2000). This corpus contains 30 files for one Spanish-speaking child between the ages of 1;7 and 2;10. We extracted the 13,411 child-directed utterances from the files using the Python package `PyLangAcq` (Lee et al., 2016). Then, we used the Python package `spaCy` (Honnibal and Montani, 2017) to parse these utterances. This allowed us to extract utterances where two clitics preceded a verb; *i.e.*, we extracted the sentences with clitic clusters that are relevant for learning the PCC. We found 50 instances of 1 3, 148 instances of 2 3, 4 instances of 3 2, and 68 instances of 3 3. This indicates that the speakers in this corpus speak a Me-First PCC language, since these constructions are only compatible with that kind of PCC language. We failed to observe any instances of 1 2, even though this construction is grammatical in Me-First PCC languages (cf. Table 1).

Training corpora for our models were created based on the frequency distribution found in the Aguirre Corpus. Because counts from this corpus were used as the weights for the random sampling, we applied smoothing so that the simulations had some probability of including the 1 2 construction, which is grammatical in Me-First PCC languages (again, cf. Table 1) but had a zero count in the Aguirre corpus. The smoothing consisted of adding 0.1 to all of the counts for grammatical constructions from the Aguirre corpus. For each simulation, we randomly sampled n PCC constructions with weights based on the smoothed frequency profile found in the Aguirre corpus; we did this for three values of n: 66, 666, and 6,666. These values were chosen because Hart and Risley (1995) estimate that children hear 333,333 utterances per year in their first three years of life. Moreover, 2% of the utterances in the Aguirre Corpus were relevant for learning the PCC, so 2% of 333,333 is 6,666 (see subsection 5.3 for more discussion).

5.2 Results

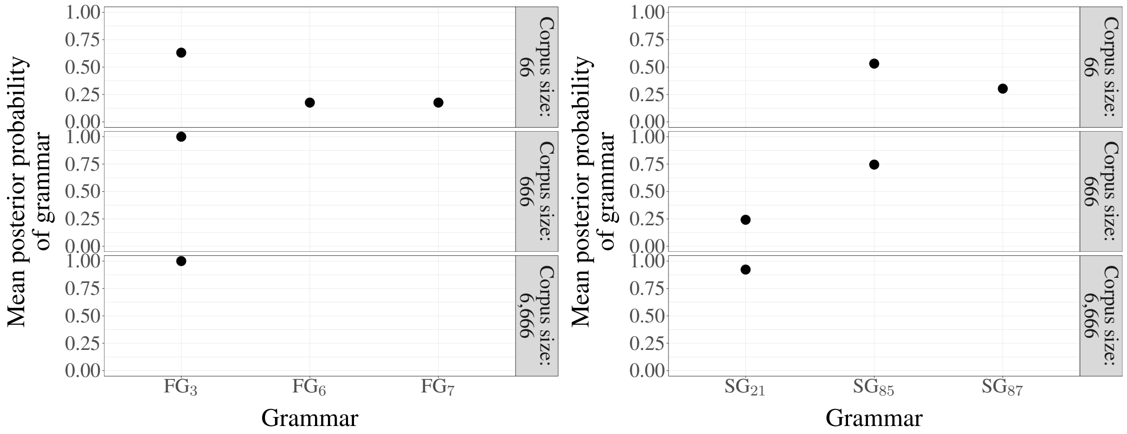
We trained Simple learning models and Feature-based learning models. Each model used the data that we generated on the basis of the Aguirre corpus to compute a posterior distribution over all the grammars in its hypothesis space. We ran 1,000 replications of each model at each corpus size, n, and we averaged the results of these 1,000 replications. These mean posterior probabilities are plotted in Figure 2 (to make the plots more readable, only grammars with a posterior probability equal to or greater than 0.1 are plotted).

As can be seen in Figure 2, the grammar with the highest posterior probability is the correct grammar for all three corpus sizes under the Feature-based learning model. That is to say, in these cases, the model has converged on FG_3, which is the feature-based grammar for the Me-First PCC (cf. fn. 6).

On the other hand, for the Simple learning model, the simulations converge on SG_{85} when the corpus size is 66 and 666, but the simple grammar that instantiates the Me-First PCC variety is in fact SG_{21}; SG_{85} differs from SG_{21} in disallowing 1 2. (SG_{87} furthermore disallows 3 2, compared to SG_{21}; see Table 2.) Nonetheless, when the corpus size is 6,666, the Simple learning model does correctly converge on SG_{21}.

5.3 Discussion

In our simulation results, we saw that the Feature-based learning model is able to converge on the correct grammar much quicker than the Simple learning model. In fact, if data are sparse, the Simple learning model converges on unattested PCC varieties. The Simple learning model clearly needs more data to learn the target grammar.



(a) Feature-based learning model results

(b) Simple learning model results

Figure 2: Mean posterior probabilities for learning simulations (FG_3 is the target grammar for the feature-based theory; SG_{21} is the target grammar for the simple theory)

As noted, we chose the corpus sizes that we did because Hart and Risley (1995) estimate that children hear 1 million utterances in their first 3 years, or 333,333 utterances per year.[7] Moreover, the Aguirre corpus contained 13,411 child-directed utterances, and we found 270 utterances with clitic clusters, which is $\approx 2\%$. Two percent of 333,333 is 6,666. Thus, a young learner might hear 6,666 clitic combinations in one of their early years of life.

This suggests that the Simple learning model may in fact have enough data that it needs in order to converge on the correct target grammar, but there are several things one would want to further investigate. First, one would want to know when a child has fully acquired the PCC restrictions of their language. To the best of our knowledge, there is very little research on this. Tsakali and Wexler (2010) reported that Greek-acquiring children seem to know the PCC restrictions of their language by age 5, but they tested this by eliciting acceptability judgments, which are often hard to do with younger children. At best, this might be an upper bound for when children know the PCC restrictions of their language. Indeed, Blasco (2000) showed that Spanish-acquiring children were correctly producing both accusative and dative clitics in Spanish by the age of 2;2, if not earlier.[8] Whether this means that they know the PCC restrictions at such a young age is an open question.

Second, there is a difference between input and intake (cf. Omaki and Lidz, 2015); that is to say, just because a learner hears 6,666 clitic clusters, does not mean that the learner uses those utterances for learning. A learner might be inattentive, a learner might fail to perceive a given utterance, a learner might fail to parse a given utterance, *etc.*. Especially at a very early age, when the child hasn't yet learned the syllable structure of their language and how to identify morpheme boundaries, it seems unlikely that the child would learn anything about the PCC variant of their target language upon hearing a clitic cluster in their input.

Moreover, as can be seen by examples (1) and (2), the surface string order does not necessarily reflect the underlying argument structure relations, which can interact with other language specific factors in a variety of ways. For example, in many dialects of Spanish, the clitics must occur in a certain order, regardless of the underlying argument structure relations (cf. fn. 1). Absent definitive knowledge of both the argument structure of the verb and such language specific factors as clitic ordering effects, it might be advantageous for a learner to ignore some of its input (cf. Perkins et al., 2017).

Thus, if a child really did know the PCC variant of their target language by age 2;2, our results might argue against the Simple learning model, if not all of the clitic clusters in the child's input are taken up and used for learning. Nevertheless, there is much we don't yet know about the acquisition of

[7] These estimates are for American children who are acquiring English, but presumably the order of magnitude is comparable for learners of other languages, such as Spanish.

[8] For further discussion on the acquisition of clitics more generally, see Tsakali (2014).

the PCC.

Additionally, there is more that could be done on the modeling side of things. For example, the models we've presented abstract away from additional complexities of the assumed grammars, such as the necessity of the Agree operation for Nevins's (2007) theory of the PCC or the necessity of the features Author and Participant. If such additional complexities also need to be learned, (*i.e.*, if they are not already known at the time when PCC learning begins), one would want to create learning models that include these complexities and run further simulations.

Ultimately, this work is intended as a computational-level analysis that begins to help set an upper bound on how much data children would need to use in order to learn the PCC, given particular theoretical and representational assumptions. We've compared the feature representations assumed by Nevins's (2007) feature-based theory to the feature representations assumed in a simple theory of the PCC. In addition to Nevins's (2007) theory, there are other more restrictive theories of the PCC (*e.g.*, Béjar and Rezac, 2003; Pancheva and Zubizarreta, 2018; Graf, 2019); so future modeling work should also seek to establish upper bounds for the theoretical and representational assumptions of these analyses. Given that they're more restrictive theories, one might expect the results to be similar to the results for the Feature-based learning models reported here, but such modeling work may nevertheless help distinguish between them, when coupled with better information about the acquisition of the PCC.

6 Conclusion

In this paper, we used a learning model to investigate how person features might be represented in the syntactic component of the grammar. We compared two possibilities: one where the person features are represented as atomic units (cf. (3)) and one where the person features are represented as feature bundles, consisting of values for the binary features Author and Participant (cf. (3)).

We simulated different-sized corpora based on realistic distributions in the input to children and evaluated these learning models against the simulated data. We found that the Feature-based learning model is able to learn the target grammar much quicker than the Simple learning model. Given enough data, the Simple learning model will converge on the correct grammar; however, if data are sparse, the Simple learning model will converge on unattested PCC variants, which might tell against the simple theory of the PCC. That is, this suggests that the larger hypothesis space, in addi-tition to being possibly unparsimonious, may lead learners astray, particularly if data are sparse.

One would particularly want to know how much input the child actually gets, how much of that the child uses, and when the child has fully acquired the PCC restrictions. Such information, coupled with our results, would inform whether one of these ideas about the representation of person features in the grammar is more plausible than another.

Acknowledgments

We gratefully acknowledge the support of NSF NRT award DGE-1449815, which has supported the first author.

References

David Adger and Jennifer Smith. 2010. Variation in agreement: A lexical feature-based approach. *Lingua*, 120(5):1109–1134.

Carmen Aguirre. 2003. Early verb development in one Spanish-speaking child. In Dagmar Bittner, Wolfgang U. Dressler, and Marianne Kilani-Schoch, editors, *Development of Verb Inflection in First Language Acquisition: A Cross-Linguistic Perspective*, number 21 in Studies on Language Acquisition, pages 1–26. Mouton de Gruyter, Berlin, Germany.

Susana Béjar and Milan Rezac. 2003. Person licensing and the derivation of PCC effects. In Ana Teresa Pérez-Leroux and Yves Roberge, editors, *Romance Linguistics: Theory and Acquisition*, number 244 in Current Issues in Linguistic Theory, pages 49–62. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Maria Blasco. 2000. *The Acquisition of Pronominal Object Clitics in Spanish*. Ph.D. thesis, The City University of New York, New York, NY.

M. Eulàlia Bonet. 1991. *Morphology after Syntax: Pronominal Clitics in Romance*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

M. Eulàlia Bonet. 1994. The person-case constraint: A morphological approach. In Heidi Harley and Colin Phillips, editors, *The Morphology-Syntax Connection*, number 22 in MIT Working Papers in Linguistics, pages 33–52. Cambridge, MA.

Hagit Borer. 1984. *Parametric Syntax: Case Studies in Semitic and Romance Languages*. Number 13 in Studies in Generative Grammar. Foris Publications, Dordrecht, The Netherlands.

Andrea Calabrese. 1995. A constraint-based theory of phonological markedness and simplification procedures. *Linguistic Inquiry*, 26(3):373–463.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht, The Netherlands.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2014. *Bayesian Data Analysis*, 3rd edition. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL.

Thomas Graf. 2012. An algebraic perspective on the person case constraint. In Thomas Graf, Denis Paperno, Anna Szabolcsi, and Jos Tellings, editors, *Theories of Everything: In Honor of Ed Keenan*, number 17 in UCLA Working Papers in Linguistics, pages 85–90.

Thomas Graf. 2019. Monotonicity as an effective theory of morphosyntactic variation. *Journal of Language Modelling*, 7(2):3–47.

Betty Hart and Todd R. Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H. Brookes Publishing Co., Inc., Baltimore, MD.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jackson L. Lee, Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. 2016. Working with chat transcripts in python. Technical Report TR-2016-02, Department of Computer Science, University of Chicago.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Andrew Nevins. 2007. The representation of third person and its consequences for person-case effects. *Natural Language & Linguistic Theory*, 25(2):273–313.

Akira Omaki and Jeffrey Lidz. 2015. Linking parser development to acquisition of syntactic knowledge. *Language Acquisition*, 22(2):158–192.

Roumyana Pancheva and Maria Luisa Zubizarreta. 2018. The person case constraint: The syntactic encoding of perspective. *Natural Language & Linguistic Theory*, 36(4):1291–1337.

Lisa Pearl, Timothy Ho, and Zephyr Detrano. 2017. An argument from acquisition: Comparing English metrical stress representations by how learnable they are from child-directed speech. *Language Acquisition*, 24(4):307–342.

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

Lisa Pearl and Jon Sprouse. 2019. The acquisition of linking theories: A tolerance principle approach to learning UTAH and rUTATH. Ms., University of California, Irvine, CA and University of Connecticut, Storrs, CT.

Laurel Perkins, Naomi H. Feldman, and Jeffrey Lidz. 2017. Learning an input filter for argument structure acquisition. In *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, pages 11–19.

David Pesetsky and Esther Torrego. 2007. The syntax of valuation and the interpretability of features. In Simin Karimi, Vida Samiian, and Wendy K. Wilkins, editors, *Phrasal and Clausal Architecture: Syntactic Derivation and Interpretation*, volume 101 of *Linguistik Aktuell/Linguistics Today*, pages 262–294. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Ezer Rasin and Roni Katzir. 2017. A learnability argument for constraints on underlying representations. Ms., Leipzig University and Tel Aviv University. Available at `https://ling.auf.net/lingbuzz/002260`.

Joshua B. Tenenbaum and Thomas L. Griffiths. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640.

Vina Tsakali. 2014. Acquisition of clitics: The state of the art. In Kleanthes K. Grohmann and Theoni Neokleous, editors, *Developments in the Acquisition of Clitics*, chapter 5, pages 161–187. Cambridge Scholars Publishing, Newcastle, UK.

Vina Tsakali and Kenneth Wexler. 2010. The acquisition of Person Case Constraint in Greek. Paper presented at the *19th International Symposium on Theoretical and Applied Linguistics*, Thessaloniki, Greece.