Learning Constructions and the Theory of Grammar

Peter W. Culicover
The Ohio State University

Andrzej Nowak
Wojciech Borkowski
University of Warsaw

1. Introduction

For a number of years we have been developing a computational simulation of language acquisition. The objective of this simulation is to explore the interactions between and the effects on the course and content of acquisition of three main factors in the theory of language acquisition: (i) the computational capacities of the learner and its ‘prior knowledge’ about language and grammar, (ii) the input to the learner, and (iii) the target of learning. The outcome of the simulation is an explicit representation of the knowledge of the simulated language learner that can be examined in detail, and compared with what human learners are presumed to know.

We have proceeded by making minimal assumptions about each of (i)-(iii). In the case of (i), we assume that the learner has no knowledge of grammatical categories, linguistic structure, or grammatical principles. The learner has only the capacity to extract correspondences between form and meaning based on the statistical properties of the linguistic input, to form categories based on similarity of distribution, and to form limited generalizations. In the case of (ii), we assume that the learner is presented only with pairs consisting of forms (including phrases and sentences) and their corresponding meanings. In the case of (iii), we assume that the target of learning is not a grammar in the sense of Mainstream Generative Grammar,[1] but a set of form-meaning correspondences that is coextensive with the form-meaning correspondences computed by a suitable grammar for the language to be learned (e.g., the grammar that is in the head of a native speaker of the language).

Clearly, these assumptions are in many respects too strong, and a realistic account of how language is acquired will have to elaborate (i)-(iii) in many ways. The objective of the simulation is not to demonstrate that the strongest form of these assumptions is correct, but to determine in exactly which ways they are too strong. Moreover, it is of some interest to discover how far a language learner can get, even given these very minimal assumptions.

We have explored these issues at some length elsewhere, both conceptually and experimentally. For example, Culicover 1999 argues that there are no inherently universal syntactic categories, and that those universals that are attested are the consequence of the projection of linguistic meaning (aka Conceptual Structure in the sense of Jackendoff 1983), onto linguistic form. Culicover 1999 argues that a very concrete grammatical representation along with Conceptual Structure is sufficient to account for the syntactic knowledge of a native speaker, for the capacity of learners to acquire grammar, and for the range of variation that is actually found in natural languages. Culicover and Nowak 2003 characterize language acquisition as the growth of a dynamical system that computes form-meaning correspondences, and report on some preliminary experiments with a computational simulation of such a system. Culicover, et al. 2005 elaborate in considerable detail the position taken by Culicover 1999 that syntactic structure is

[1] We use this term as a convenient way to refer to the approach found in syntactic theories from Chomsky 1957 through Chomsky 1994, which although varying in specifics over the years, share certain features (see Jackendoff 2002 Culicover, et al. 2005). The reader is welcome to substitute whatever term s/he prefers for this purpose.

concrete and that much of the burden of explaining how language works can be carried by Conceptual Structure and the syntax-semantics interface (the "Simple Syntax Hypothesis").

There is a very important consequence of this perspective that is particularly relevant to the topic of this Forum. It is that the learner is assumed to start language acquisition by treating every form-meaning correspondence as a 'construction,' in the sense that it is more specific and idiosyncratic than a general rule of grammar. For instance, English has the general phrase structure rule *VP → V …* . It is reasonable to think that adult speakers of English have some form of this rule as part of their knowledge of the language. But in the simulation, a learner will first learn that *eat (your) spinach* corresponds to a particular meaning (e.g. EAT($THEME:SPINACH), that *kiss (the bunny)* corresponds to a particular meaning (e.g. KISS($THEME:BUNNY)), and so on. It is only after some time and experience that the learner begins to form generalizations that approximate *VP → V …* . If the correspondence given to the learner is less general, the correspondence formulated by the learner will be less general.

The alternative view, which the work cited above argues against, is that syntactic structures such as that of VP, and other particular aspects of syntactic knowledge, are available to the learner from the outset, and allow variation only within the set of possibilities permitted by 'parameters.' Conceded, there may in fact be certain aspects of linguistic knowledge that the learner must have in place in order to be able to acquire a language; it is one of the goals of the simulation to discover what they are. But there is nothing in this simulation akin to parameter setting, since by assumption there is no prior hypothesis space, and no parameters.[2]

2. Construction(s): The Correspondence Spectrum

A question that now naturally arises is to what extent 'construction' plays a role in the characterization of knowledge of language in the adult. If plays a central role, then it is natural to envisage a continuity of development, wherein numerous primitive constructions are formed at the earliest stage of learning and gradually coalesce and generalize as the learner is exposed to more and more convergent linguistic input. On this view, general grammatical rules of the familiar sort would be the limiting cases, but not the only components of knowledge of language.

Pursuing this idea, we adopt the Jackendoffian perspective on grammar (see, e.g., Culicover & Jackendoff, *Syntax Made Simple(r),* Oxford University Press, to appear), which is a constructionalist one with a vengeance. Its main features are these:

a. The job of grammar is to describe the form-meaning correspondences .

b. **Simple(r) Syntax Hypothesis (SSH)**: The most explanatory syntactic theory is one that imputes the minimum structure necessary to mediate between form and meaning.

c. Some of the correspondences are unanalyzable (words).

d. Some have structure but are simple or not transparent on the meaning side (idioms) (no structure/meaning match-ups).

e. Some have structure and are transparent on the meaning side (compositional semantics interpreting canonical phrase structure).

f. Some are a combination of the above ('constructions'), ranging from quasi-idioms, double-objects, movement along a path, syntactic nuts (Culicover 1999), various operator-trace binding constructions such as wh-questions, topicalization, etc. Each has some degree of predictability and generality, some idiosyncrasies.

[2] Our approach is consistent with the "constructionist manifesto" of Quartz and Sejnowski 1997: "In contrast to learning as selective induction, the central component of the constructivist model is that it does not involve a search through an a priori defined hypothesis space, and so is not an instance of model-based estimation, or parametric regression. Instead, the constructivist learner builds this hypothesis as a process of activity-dependent construction of the representations that underlie mature skills."

We briefly review several types of correspondence mentioned here and suggest that a constructionalist perspective is preferred to derivational alternatives.

2.1. Words

Some linguists (e.g. Hale and Keyser 1993) have argued that apparently simple words are syntactically complex and are the product of derivations involving movement and deletion. But the relations captured by such derivations can be captured in non-derivational (constructionist) ways, and the latter are required anyway for certain aspects of the correspondences. For example, the verb *shelve* is morphologically and semantically related to the noun *shelf*; it means, roughly, ‘put on (a) shelf’. But the verb has idiosyncratic phonological and semantic properties that cannot be predicted from the noun alone. For example, the form is *to shelve,* not **to shelf,* and use of the verb is restricted to things like books that are commonly put on shelves. Hence we cannot say **I shelved the cat* to mean ‘I put the cat on the shelf’ except as a joke. Hence it is necessary to associate with the word *shelve* its particular form and meaning, an individual lexical correspondence, and this is sufficient to represent its relationship to the noun *shelf.*

2.2. Idioms

Some idioms have quasi-transparent interpretations and regular syntax. Others are essentially opaque. The following examples occupy various positions on the scale of transparency.

(1) by and large
lo and behold
beat a dead horse
make amends
cast aspersions
a flash in the pan

In each case (and of course there are vastly more that we haven’t listed), it is necessary to specify details of form and interpretation. Each of these idioms is an individual construction, somewhat more complex than a word, but much more specific than a phrase structure rule.

2.3. VP constructions

Next we come to idioms that are composed of some general (i.e. categorical) requirements and some specific lexical requirements.

(1) a. *Pat sang/drank/sewed his heart out*. [also *his guts*]
b. *Terry yelled/wrote/programmed her head off*. [also *her butt, her tush*, etc.]
c. *Leslie talked/cooked/composed up a storm.* [**Leslie talked a storm up.*]

Normal verb-particle constructions:
d. *Pat threw the trash out.*
e. *Leslie picked up the garbage.*

An idiom such as *V one’s N out/off* is semantically intransitive but syntactically transitive. Because of the latter, it cannot be used transitively, i.e. with a THEME argument.

(2) a. **Pat sang the Marseillaise his heart out.*
b. **Terry yelled insults her head off.*
c. **Leslie cooked eggs up a storm.*

Here are some other VP idioms.

(3) a. *Way*-construction (Jackendoff 1990; Goldberg 1995):
Elmer hobbled/laughed/joked his way to the bank.
('Elmer went/made his way to the bank hobbling/laughing /joking')
b. Time-*away* construction (Jackendoff 1997b):
Hermione slept/drank/sewed/programmed three whole evenings away.
('Hermione spent three whole evenings sleeping/drinking/sewing /programming')
c. Sound+motion construction (Levin and Rappaport Hovav 1995):
The car whizzed/rumbled/squealed past Harry.
('the car went past Harry, making whizzing/rumbling/squealing noises')
d. Resultative construction
The chef cooked the pot black.
('the chef made the pot black by cooking in/with it')

The approach urged by Goldberg 1995 and Jackendoff 1997 (also Goldberg & Jackendoff 2004) is to view the constructions in (3), like (1a-c), as lexical VP idioms with open verb positions. Unlike (1a-c), these idioms also select other arguments - within VP to be sure, but not selected by the verb:

(4) a. [VP *V X's* way *PP*], 'go PP, while/by V-ing'
b. [VP *V NP* away], 'spend [NP amount of time] V-ing'
c. [VP *V PP*], 'go PP, making V-ing noise as a result of motion'
d. [VP *V NP AP/PP*], 'make NP become AP/PP, by V-ing'

As in the case of *V one's N out/off,* because the idiom dictates the form of the VP, there is no room for the verb to have its own arguments there; this is why the verb must be intransitive.

2.4. Syntax-semantics mismatches

These examples illustrate the fact that the form-meaning correspondence in general is not stateable simply in terms of lexical and phrasal categories. There are idiosyncrasies up and down the line, in terms of the form and in terms of the meaning. All of these constructions share the same basic syntax (not surprisingly, since they are all English); what is idiosyncratic is the way in which their meanings are related to the meanings of the parts and the structure in which they (the parts) appear. They are in the middle of a spectrum that extends from individual words at one end to structures whose form and meaning can be characterized in general terms. These general cases seem to be transparent, that is, compositional and non-'transformational'.

(2) a. *eat the bagel* EAT(x,BAGEL)
b. *see a dog* SEE(x,DOG)

However, while V-NP may be relatively transparent, even simple V [$_{NP}$ Adj N] gets complicated.

(3) *eat a quick/occasional/leisurely/?slow bagel*
*eat the *quick/occasional/*leisurely/*slow bagel*
smoke a quick/occasional/leisurely/ ?slow cigar

(4) *?bought a quick/occasional/leisurely/?slow bagel* (ok w/intention to consume)
*burned a *quick/occasional/*leisurely/*slow bagel*
see an/the occasional dog

Not to mention $[_S$ NP-VP]

(5) *The occasional dog attacked me on my run.*

Mismatches are very common and are part of what needs to be learned.

3. *CAMiLLE*

Thus, the facts of a language such as English suggest that what is to be learned includes a large set of correspondences, or constructions, ranging in generality from individual words to compositional phrase structures. Since there is no way for the learner to know where on the spectrum a correspondence really is, the conservative strategy is to start at the word/idiom end, and then move away as the weight of the evidence warrants generalization (Tomasello 2000). The weight of the evidence is at least in part statistical (e.g. Newport & Aslin 2004). As noted earlier, the question that we are concerned with in our simulation is what other factors, if any, have to be brought into play in order to account for the learner's ability to achieve the target on the basis of linguistic experience.

Our approach to this question is one of 'Concrete Minimalism' (Culicover 1999). We assume that the computational system is maximally simple, not in terms of abstract computational simplicity (as in Chomsky 1994), but in terms of the criterion of learning on the basis of the concrete evidence. That is, it should be the simplest system that can arrive at an adequate account of the language given a large but finite sample of experience.

The simulation is called ***CAMiLLe***:

C	onservative (or Concrete) (don't generalize much beyond the evidence)
A	ttentive (all input is potentially relevant)
Mi	nimalist
L	anguage
Le	arner

Pursuing the logic of Concrete Minimalism, we constructed *CAMiLLe* with minimal prior knowledge of linguistic structure. Language acquisition by *CAMiLLe* is intended to simulate the formation of trajectories and flows, and self-organization, in a dynamical system. Our experiments with *CAMiLLe* are intended to determine how much grammatical knowledge such a minimalist learner is capable of acquiring strictly from sound/meaning pairings.

3.1. What CAMiLLE does

CAMiLLE is exposed to sets of form-meaning pairs, e.g.

(6)
```
house = HOUSE
see the house ? = YNQ(SEE($EXP:YOU,$THEME:HOUSE))
```

On the basis of collections of such pairs, ***CAMiLLE*** attempts to formulate correspondence rules. The rules are formulated by grouping sentences whose meaning contain a particular meaning component into one, and grouping those whose meaning lacks this meaning component into another set. ***CAMiLLE*** then formulates a hypothesis with a certain probability attached to it that particular features of the sentence correspond to the meaning component. It does this for all meaning components, and may hypothesize several rules at the same time.

It is important to point out that ***CAMiLLe*** proceeds from the assumption that strings of words and their corresponding meanings are organized according to heads and non-heads (dependents or adjuncts). Since there is no overt connection between the individual words and the individual meanings, ***CAMiLLE*** is prone to making many bad rules. But ***CAMiLLE*** also will make correct rules. For example, after having encountered the sentences in (7) –

```
(7)   ted is nice. = BE($THEME:TED, $PRED:NICE)
      ted is small. = BE($THEME:TED, $PRED:SMALL)
```

– ***CAMiLLE*** has enough information to guess that *Ted is* means either `BE($THEME:TED)`, `TED`, or `BE`. ***CAMiLLE*** keeps track of the evidence that supports each hypothesis, so that after enough experience, the diversity of exemplified correspondences continues to suppose the first hypothesis, but not the other two. At the same time, this experience provides evidence that *Ted* corresponds to `TED` and *is* corresponds to `BE.` The evidence is purely statistical; the rules that are not supported remain but gradually get pushed out by rules that are more strongly supported by the evidence.

If ***CAMiLLE*** finds that two rules have a similar form, then to the extent possible it forms a cluster (i.e. a mini-category). For example, if ***CAMiLLE*** has strong evidence for the following two correspondences --

```
(8)   ted is <=>  BE($THEME:TED)
      sally is <=>  BE($THEME:SALLY)
```

Then ***CAMiLLE*** will form a correspondence rule of the form

```
(9)   [ted;sally] is <=> BE($THEME:[TED;SALLY])
```

Clearly, correspondences such as these are not equivalent to rules of grammar in the traditional sense. For one thing, they are much too specific – they do not mention categories but simply clusters of individual elements. For another, they provide information only about the linear order of elements, not structure. And they do not provide any phrasal information.

At the same time, it is possible that what ***CAMiLLE*** comes up with is comparable in some important respects to what an early language learner comes up with, prior to the point at which generalization and the formation of large-scale categories and correspondences kicks in. We are entertaining the hypothesis (suggested by Tomasello 2003 and arrived at independently through our own preliminary experiments with ***CAMiLLE***) that first there is a pre-grammatical stage, which is modeled by ***CAMiLLE***, followed by a grammatical stage. In the pre-grammatical stage we expect to see everything treated as though it is a construction. In the grammatical stage, we expect to see those aspects of the language that are fully or almost fully regular to be reflected in dramatic generalizations, while those aspects of the language that retain some significant idiosyncrasy, e.g. constructions of the sort that we noted in §2, would be retained in their pre-grammatical form.

The next two sections summarize some of the results of experiments with ***CAMiLLE***. In §3.2 we discuss some experiments with constructed input, and in §3.3 experiments with naturally occurring input.

3.2. Constructed input

Constructued input allows us to test ***CAMiLLE***'s ability to deal with a particular grammatical phenomenon. ***CAMiLLE*** requires a certain amount of exposure to a grammatical phenomenon in order to form a reasonably informed hypothesis about it. A file of naturally occurring, transcribed speech to children from the CHILDES database (MacWhinney 1995) in general does not provide enough instances of a specific phenomenon,[3] and running ***CAMiLLE*** on composites of files, while potentially useful (see §3.3) does not allow us to focus on specific grammatical phenomena. So we have constructed files. An example of a constructed input file is given below as **Sample Input 1: word_order-1.txt.**

Sample Input 1: word_order-1.txt

```
house = HOUSE
see the house ? = YNQ(SEE($EXP:YOU,$THEME:HOUSE))
mary = MARY
here's mary = $POINT($THEME:MARY)
see mary ? = YNQ(SEE($EXP:YOU,$THEME:MARY))
john = JOHN
see john ? = YNQ(SEE($EXP:YOU,$THEME:JOHN))
here 's a flower = $POINT($THEME:FLOWER)
see the flower = $IMP(SEE($EXP:YOU,$THEME:FLOWER))
here's a boy = $POINT($THEME:BOY)
see the boy = $IMP(SEE($EXP:YOU,$THEME:BOY))
horsie = HORSE
see horsie ? = YNQ(SEE($EXP:YOU,$THEME:HORSE))
look, a baby ! = BABY
see the little baby ? = YNQ(SEE($EXP:YOU,$THEME:BABY))
nice baby ! = BABY
talk to the baby = $IMP(TALK($AGENT:YOU,$GOAL:BABY))
talk to mary = $IMP(TALK($AGENT:YOU;$GOAL:MARY))
i see mary = SEE($EXP:ME,$THEME:MARY)
i am talk ~ing to mary = TALK($AGENT:$ME;$GOAL:MARY)
do you see john ? = YNQ(SEE($EXP:$YOU,$THEME:JOHN))
```

The purpose of this particular file is to try to get ***CAMiLLE*** to correlate individual words with their meanings, and to correlate position in the string with semantic role. (The roles used here are THEME, EXP(ERIENCER) and AGENT.) The output after processing ten sentences consists of 103 rules, many of them overlapping, and many of them highly idiosyncratic but low in weight. Lack of space precludes listing all of the rules here, so we will show a few 'correct' rules (10) and a few 'incorrect rules' (11).

(10)

```
5.  [89]   MARY <=> mary
6.  [82]   JOHN <=> john
16.  [13]   YNQ(*NULL*:SEE) <=>   1.see   3.?
21.  [10]   $IMP(*NULL*:SEE <=> 1.see
27.  [9]   SEE($THEME:[JOHN; MARY;]) <=> see+1->[john; mary;]
95.  [2]   $POINT($THEME:MARY) <=>  1.here's  2.mary
```

(11)

```
1.  [172] SEE($EXP:YOU) <=> 1.see
23.  [10] YNQ(*NULL*:SEE)$=#3   <=> 1.see  2.the  3.?
72.  [2]  BOY <=> 3.boy  |
```

[3] It is of course an empirical question whether for any given grammatical phenomenon, the naturally occurring data taken as a whole provides sufficient evidence for a learner. If it does not, then this is an argument (from poverty of the stimulus) for innateness. The sorts of things that we are interested in are those that are not universally found in languages of the world, and therefore we may pretty confidently presume that they are learned on the basis of evidence in the linguistic input to the learner.

```
                   a+1->boy see  |
                   see X boy
103. [2]  FLOWER <=>   1.here  2.'s  3.a  4.flower  |
                       1.see  2.the  3.flower
```

The ‘correct’ rules and the ‘incorrect rules’ all reflect ***CAMiLLE***’s exposure to the data. For instance, (11.1) reflects the fact that *see* in initial position correlates highly with the meaning SEE($EXP:YOU). This is because there are a lot of sentences beginning with *see* (imperatives and questions) in which the subject is not expressed overtly. We may take this to be a very early stage of development, in which the learner has not yet determined that such sentences have a missing subject; such a determination can be made when the learner recognizes that all sentences of English have subjects.

More strikingly, (11.23) shows that in the limited input data, *see the* correlates highly with the interrogative of SEE. This is an artifact of the particular dataset, and is not an error on ***CAMiLLE***’s part, but a correct hypothesis under the circumstances. Similarly, ***CAMiLLE*** finds evidence to form correspondences between the meaning BOY and *boy* in third position, *a boy*, and *see ... boy*. A more diverse set of experiences will disabuse ***CAMiLLE*** of these errors. And it is possible, although difficult to determine experimentally, that actual learners may form such incorrect, yet fleeting, mistaken correlations in the early stages of leanrning.

Rule (10.16) reflects the fact that *see* is used as an interrogative (with ‘?’ in third position in the sentence – an artifact of the input data). Similarly, (10.21) reflects the interrogative case. (10.27) indicate that ***CAMiLLE*** has identified *John* and *Mary* as elements that have the same distribution (with respect to the THEME of *see*). This observation may, if we wish, form the basis for a generalization that *John* and *Mary* have the same distribution with respect to everything, although we will want to exercise caution in formulating the rule of generalization. Finally, (10.95) is a small construction, correlating *here’s Mary* with pointing to Mary.

3.3. Natural Input

We have begun to look at what ***CAMiLLE*** does with natural linguistic input. There are approximately 775K English sentences spoken to children in the CHILDES database. As we have mentioned, ***CAMiLLE*** learns by determining correspondence rules mapping form and meaning. Since the sentences in the CHILDES database do not have meanings associated with them, in order to use those sentences as input to ***CAMiLLE*** it is necessary to provide them all with meanings. Doing so manually is prohibitively labor intensive.[4] Our approach has been to parse the sentences with a fast (but, unfortunately, inaccurate) parser (Mini-par), translate the output of the parser into rudimentary meanings, present the resulting sets of sentence/meaning pairs to ***CAMiLLE*** and ask ***CAMiLLe*** to figure out the correspondence rules.

The results thus far are somewhat inconclusive, but for reasons that do not necessarily reflect on ***CAMiLLE***. First, the 775K sentences in the CHILDES database are taken from many speakers, a number of dialects, and are spoken to children of widely varying ages. This means that there may be an overall lack of consistency that might interfere with ***CAMiLLE***’s ability to extract reasonably accurate generalizations. Second, and more seriously, the parses produced by Mini-par are often wildly mistaken; hence the meaning that is automatically generated from the parse is also wildly mistaken. To take just one example, Mini-par treats *going to* as a directional, even when it is used in sentences like *Are you going to kiss me?* Conjunction, which is found very often in the CHILDES data, is also very problematic for Mini-par (and all other parsers, for that matter). Thus, there are numerous errors that arise out of the misparsing of the input. Finally,

[4] If it takes one minute to construct each meaning, and a person does this eight hours a day, five days a week, it would take over six years to assign meanings to each of these sentences. Who knows how long it would take to correct the errors.

the range of subject matters found in the CHILDES data is quite extensive, and a successful meaning assignment to a large number of sentences, even if they are correctly parsed, is a non-trivial task that we have yet to contemplate.

This being said, ***CAMiLLE*** does produce some useful output when dealing with the naturally occurring input, and some of this output very definitely has the feel of preliminary constructions. The following rules are a small sample of what ***CAMiLLE*** came up with after processing approximately 19K sentences from the CHILDES database.

(12)

```
1. [59152] BE($THEME:[HE; IT; THAT; THIS; WHAT; WHERE; WHO;]) <=> [he; it;
that; this; what; where; who;]+1->is
6. [38487] [BED; BOOK; BUG; BUNNY; CHAIR; COOKIE; CRAYON; DUCK; IT; PICTURE;
THAT; THIS;]($REF:[$DEF; $INDEF;]) <=> [a; the;]  [?; bed; book; bug; bunny;
cookie; duck; picture;]
15. [12036] [KNOW; LIKE; SEE; THINK; WANT;]($EXPERIENCER:YOU) <=> you  [know;
like; see; think; want;]
62. [1491] IMP([GET; GIVE; LIKE; PUT; SAY; SEE;]) <=> 1.[get; give; like; put;
say; see;]
155. [372] [WHAT; WHERE;]($REF:$WH) <=> 1.[what; where;]  is  |
                                         [what; where;]+1->is
163. [353] PLAY_WITH <=> play+1->with
195. [266] YNQ(*NULL*:WANT) <=> do X want
228. [227] WANT <=> want+1->to
590. [43] NEG(*NULL*:BE) <=> 2.is  3.not
```

Rule (12.1) characterizes a construction in which there is a pronominal THEME of *be*. (12.6) shows that *a* and *the* correlate with the features DEF and INDEF on nominal concepts. This may ultimately form the basis for a more general rule NP → Det-N, although getting there requires additional generalization. (12.15) correlates *you* when it precedes a set of verbs expressing knowledge and perception with the EXPERIENCER role of these verbs, corresponding to the subject of these verbs. And so on for the rest of the rules shown. Each one indicates that ***CAMiLLE*** has correctly extracted some correspondence, one that is specific to lexical items. This is how we characterized the most specific end of the correspondence spectrum in §2. Hence ***CAMiLLE*** appears to be capable, in principle at least, at carrying out the preliminary work of forming correspondence rules.

4. Summary and prospects

There is clearly a lot more that can be said about what ***CAMiLLE*** does, if only because even in its current form, it produces so much output. The massive output provided by ***CAMiLLE*** is both a curse and a blessing. It is a curse, because it is so much to deal with, and not particularly easy to analyze. But it is a blessing, because what we are doing in creating ***CAMiLLE*** is simulating what takes place in the mind of a language learner. If it is fact true that early language learners begin by creating numerous constructions and only later generalize over and perhaps beyond them, then looking at ***CAMiLLE***'s output is like looking directly into the language faculty.

Of course it would be a serious mistake to claim that this program is anything more than a simulation, or that it is necessarily a correct simulation of how learning proceeds. The ultimate test will be whether ***CAMiLLE***, or a subsequent development of ***CAMiLLE***, is capable of producing a representation of the language learned that comprises in a satisfactory way a native speaker's knowledge of language (or at least, the form-meaning correspondences). Such a representation has to go beyond the actual experience. Moreover, it must capture generalizations that are formulated at a level of abstraction that goes well beyond what is available to ***CAMiLLE*** at this point.

These issues are the focus of our current work with ***CAMiLLE***. Our immediate goal, besides improving the input to the simulation, is to provide ***CAMiLLE*** with the capacity to generalize beyond individual or clustered correspondences (the outcome of the pre-grammatical stage) to correspondences in terms of general categories (the grammatical stage). We are also experimenting with various 'local' relations, such as Subject-Aux inversion, to show that ***CAMiLLe*** can master them without elaborate knowledge of syntactic structure beyond linear order. Our experiments with wh-questions and topicalization are intended to show that ***CAMiLLe*** can construct adequate local variants of these unbounded dependencies, which may serve well enough in the pre-grammatical stage.

Beyond this, it is clear that ***CAMiLLE*** is not able to identify the locus of a 'gap' in a sentence. That is, ***CAMiLLE*** cannot connect a 'moved' constituent with the corresponding canonical position. While it is likely that this capacity does not exist in early learning (see Tomasello 2000), it is something that ***CAMiLLE*** needs to be able to do at some point in the course of development. We see no way for ***CAMiLLE*** to discover that such connections exist unless ***CAMiLLE*** is endowed with the capacity to determine that something is absent from a particular position. True generalizations (i.e. those that speakers really make use of to assign interpretations to sentences and to judge acceptability) that crucially rely on grammatical notions such as SUBJECT and OBJECT, or thematic hierarchies, are also beyond the scope of ***CAMiLLE***, and would have to be built in – we see no way for ***CAMiLLE*** to discover them given just primary linguistic data.

In sum, we are able to demonstrate the feasibility of simulating the pre-grammatical stage of language acquisition; simulating the grammatical stage is yet to be done. The key to both, we believe, is the formulation, clustering and subsequent generalization of correspondences that embody constructions.

References

Chomsky, Noam. (1957). *Syntactic Structures*. The Hague: Mouton.

Culicover, Peter W. (1999). *Syntactic Nuts*. Oxford: Oxford University Press.

Culicover, Peter W., & Ray Jackendoff. (2005). *Syntax Made Simpler*. Oxford: Oxford University Press. In press.

Culicover, Peter W., & Andrzej Nowak. (2003). *Dynamical Grammar*. Oxford: Oxford University Press.

Goldberg, Adele E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, Adele E., & Ray Jackendoff (2004). 'The English Resultative as a Family of Constructions.' *Language.* To appear.

Hale, Kenneth, & Samuel Jay Keyser (1993). 'On argument structure and and the lexical expression of syntactic relations.' In Kenneth Hale & Samuel Jay Keyser, eds. *The View from building 20 : essays in linguistics in honor of Sylvain Bromberger*. Cambridge, MA: MIT Press.

Jackendoff, Ray. (1983). *Semantics and Cognition*. Cambridge: MIT Press.

Jackendoff, Ray. (1990). *Semantic Structures*. Cambridge, MA.: MIT Press.

Jackendoff, Ray (1997). 'Twistin' the Night Away.' *Language* 73:534-59.

Jackendoff, Ray. (2002). *Foundations of Language*. Oxford: Oxford University Press.

Levin, Beth, & Malka Rappaport Hovav. (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge, MA: MIT Press.

MacWhinney, Brian. (1995). *The CHILDES project : tools for analyzing talk*. Hillsdale, N.J.: Erlbaum.

Newport, Elissa L., & Richard N. Aslin. (2004). 'Learning at a distance: I. Statistical learning of non-adjacent dependencies.' *Cognitive Psychology,* 48:127-62.

Quartz, S. R., & T. J. Sejnowski. (1997). 'The Neural Basis of Cognitive Development: A Constructionist Manifesto.' *Behavioural and Brain Sciences* 20(4):537 -596.

Tomasello, Michael. (2000). 'Do young children have adult syntax.' *Cognition* 74:209-53.

Tomasello, Michael. (2003). *Constructing a Language*. Cambridge, MA: Harvard University Press.