A Computational Basis for Brown's Results on Morpheme Order Acquisition

Sheldon Nicholl and **David C. Wilkins**
Dept. of Computer Science
University of Illinois
1304 W. Springfield Ave.
Urbana, IL 61801

Abstract

This paper presents the result that a computer program can mimic the acquisition by children of a selected set of grammatical morphemes. Roger Brown [Brown, 1973] studied the acquisition of 14 morphemes, and showed how a set of partial order relations describes this aspect of child language learning. We show that these relations can be given a computational basis. They follow directly from a class of Boolean learning algorithms which have three simple constraints in the manner in which they consider hypotheses. I will call these three constraints the CAM constraints. CAM constraint 1 is to increase the length of the conjuncts one term at a time. The second CAM constraint is to consider all hypotheses of the same length simultaneously. Finally, CAM constraint 3 is to collect all single-term hypotheses involving noun features into a single conjunction prior to Boolean learning.

Introduction

The problem of how language is acquired is still one of the core problems in understanding cognition. Contributions to an understanding of language acquisition have come not only from Psychology and Linguistics [Brown, 1973], [Pinker, 1984] but also from Computer Science [Gold, 1967] and artificial intelligence [Berwick, 1985]. Therefore we would expect further progress in the study of language acquisition to draw more and more from multiple fields in the cognitive sciences, rather than to draw from purely one area.

Brown's results [Brown, 1973] provide a fine example of this interdisciplinary transfer. In studying the language development of three children, Brown focussed his attention on the acquisition of 14 particular grammatical morphemes. Brown was able to write down the acquisition order of a significant subset of these grammatical morphemes in terms of partial orders. This still stands as a major result in the study of language acquisition.

The goal of this paper is to exhibit a computer model of some of Brown's acquisition order results (see Table 3). This computer model is based on a set of constraints for whose plausibility we hope to argue. I will call these constraints the CAM constraints. The CAM constraints are part of a larger model of language acquisition called CAM (Categories, Agreement, and Morphology) which has been implemented in Common Lisp.

Morpheme	Meaning
Present Progressive	Temporary duration;
Plural	Number
Past Irregular	Earlierness
Third Person regular	Number; Earlierness
Uncontractible Auxiliary	Temporary duration; number; earlierness; (process-state)

Table 1: Excerpt from Brown's Table 60, p. 369 (Brown, 1973)

Brown's Results

Brown used the idea of a complexity ordering to lend an underlying paradigm to his order-of-acquisition data. Different grammatical constructions are ranked on a scale of increasing complexity; the more complex the construction, the longer it takes to learn. Brown considered two complexity orderings, one based on syntax, the other based on semantics. The syntax-based ordering is stated in terms of a theory which is seriously out of date (Chomsky's Standard Theory [Chomsky, 1965]) and also depends on transformations, which CAM does not represent; this ordering will therefore not be further considered here.

What is of interest here is Brown's semantic complexity ordering. It is appropriate to begin with a description of the actual semantics that form the basis of the ordering.

Table 1 shows the meanings of the plural and past morphemes.[1] Note that although Table 1 refers to both

[1] As Brown points out, the subjunctive use of the past,

regular and irregular forms, this distinction is not of consequence here, since both forms encode the same meaning. This is true of the past forms as well as the third person forms. Now the third person forms refer to both Number and Earlierness, since this morpheme can only be used in singular, present tense contexts.[2] Finally, the auxiliary refers to three features: temporary duration,[3] Number, and Earlierness. The temporary duration feature arises from the correlation between the auxiliaries with the progressive, at least in the data Brown considered:

I	am	walk	-ing
1S	AUX	V	progressive

The other features reflect the particular choice of the auxiliary: *is*, *are*, *was*, and *were* each make different commitments to Number and Tense; hence the need for the features.

Now that the meanings of the morphemes have been explained, Brown's partial orders based on those meanings can be described. Here is Brown's actual listing of the partial orders, redrawn slightly to fit in this column:

$$\left.\begin{array}{l}\text{plural}(x)\\ \text{past irregular}(y)\end{array}\right\} < \begin{array}{l}\text{Uncont. copula}(x+y)\\ \text{3rd-person reg.}(x+y)\end{array}$$

$$\left.\begin{array}{l}\text{Uncont. copula}(x+y)\\ \text{3rd-person reg.}(x+y)\end{array}\right\} < \text{Un. Aux}(x+y+z)$$

$$\text{Progressive}(z) < \text{Uncontractible Aux}(x+y+z)$$

Table 2: Brown's Table 61: "A partial ordering in terms of cumulative semantic complexity."

In a formal sense, the less-than symbol '<' means "less complex than". So for example, plural(x) < third-person regular($x+y$), a relation implied by Table 2, means that the plural, which depends on only one variable, x (number), is strictly less complex than the third person regular, which depends on two variables, x and y, where x is number and y is earlierness. The plus sign '+' just means "in some combination with"; it has no relation to arithmetic addition.

Fortunately, the purely formal, "less complex than" sense of '<' corresponds to an important real-world sense of '<': "acquired before". This means, for instance, that noun plural forms like *boys* are acquired before third-person regular forms like *John walks*. Now "acquired" in Brown's usage refers to the first speech sample that is "the first speech sample of three, such that in all three the inflection is supplied in at least 90 percent of the contexts in which it is clearly required." (Brown quoting Cazden, p. 258).[4] For most of the children a speech sample consisted of two hours of taped speech every second week (Brown, p. 52). The predictions of Table 5 are largely confirmed (Brown, p. 371) according to the 90% criterion. The goal of this paper is to provide a computational model of these results. The specific results addressed by this paper are the subject of the next section.

CAM

CAM is a semantics-based program in that it gets not only a linguistic string as input, but a representation of the string's meaning as well. In the interest of greatest generality, the representational power of the meaning formalism has been restricted as much as seems possible. Only words referring to clearly perceptible physical objects, events involving physical objects, or states involving physical objects are labelled as such in the meaning representation; all other words in the string regardless of their syntactic category are given a null lexical meaning. CAM's learning is interesting because it can still form a grammar despite the impoverished nature of its input.

As mentioned before, CAM is a language acquisition program concentrating primarily on Categories, Agreement, and Morphology. CAM has separate modules for Category learning, e.g., see [Nicholl and Wilkins, 1990], CFG rule learning, affix order learning, and agreement rule learning, among others. It is the agreement rule learner that is of primary interest in this paper, since the agreement rule learner is implemented with a simple bottom-up nominal-feature Boolean learning algorithm that obeys the three CAM constraints.

The Set of Morphemes Studied Here

For a variety of reasons I will explain shortly, CAM cannot at present acquire all 14 of Brown's morphemes. Fortunately, the CAM constraints are not affected by any of these considerations, so it is entirely possible that with appropriate refinements in other parts of CAM, all 14 morphemes may someday be covered.

A quick examination of Table 1 shows that Brown makes a distinction between the contractible and uncontractible forms of a morpheme. For example, *you're walking* is a contracted version of *you are walking*, but *are you walking?* allows no contraction, and is hence uncontractible. Although syntactic factors do have a bearing on contractibility, it is also clear that phonological factors are very important. CAM makes the important but certainly not unprecedented distinction between syntax and phonology and hence has no

which is used to refer to unactualized events, need not be considered here, since the children learn this meaning of the past long after the morphemes of interest here have already been learned.

[2]Brown has apparently suppressed Person from his Tables.

[3]I am suppressing "process-state" from my exposition because there are too many exceptions to the notion that the progressive somehow encodes the process-state distinction.

[4]This is the "90% criterion" which I will refer to later.

phonological component as yet. The fact that CAM can still duplicate some (but not all) of Brown's major results, i.e., Table 3 (but not Table 2), *is* unprecedented, as far as I know, and is the major result of this paper. The clear implication is that syntax and semantics are the principal (but not total) determinants of acquisition order in this domain.

Similarly, Table 1 makes a distinction between regular and irregular forms. This distinction is largely lexical. Since CAM focusses on the learning of general rules, irregular forms are neglected. This is borne out in the data to some extent, since the learning of irregular forms is different from that of regular forms. Brown also distinguishes between copulas and auxiliaries. CAM has no provision for copulas as such, due to the semantic indistinguishability of stative verbs and adjectives: both are states; both are predicates. But since CAM can still learn the full agreement rules for the auxiliary versions of all the copulas, this may not be much of a problem, especially if the acquisition of copulas is regarded as "parasitic" on the acquisition of auxiliaries. Possessives are excluded from the present implementation of CAM due again to the difficulty of learning states. For an enlightening discussion of the ambiguous representation of states in natural language, see [Lyons, 1977]. Finally, the prepositions *in* and *on*, like all other prepositions, fail the Relevance Property [Nicholl and Wilkins, 1990] for frequently grammaticized features, and are therefore deliberately ignored by CAM.

$$\left.\begin{array}{l}\text{plural}(x)\\ \text{past regular}(y)\end{array}\right\} < \text{Third-person regular}(x+y)$$

$$\text{3rd-person reg.}(x+y) < \text{Uncont. Aux}(x+y+z)$$

$$\text{Progressive}(z) < \text{Uncontractible Aux}(x+y+z)$$

Table 3: Order of child language acquisition experimentally observed by Brown, a subset derived from his table 61, p. 370 (also Table 2, this article). The '<' sign is Brown's notation for "acquired before." The relations here in Table 3 have been computationally duplicated by using the CAM constraints. For examples, see Table 5.

Of Brown's original 14 morphemes, this leaves the following morphemes learnable by CAM: the progressive, plural, past, third-person regular, and auxiliary verb morphemes. Fortunately, these are almost exactly the same morphemes covered by Brown's acquisition order results (Table 2) so little is lost.

The forms of the morphemes learnable by CAM are as follows. For the progressive, *-ing* is the form learned. For the plural, *-s* is the allowable form because irregular plurals are excluded. The same is true for the past: *-ed* is learned, while the irregular past forms are neglected here. As one might expect, *-s* is learned for the third-person regular. The case of auxiliary verbs is more interesting. CAM learns *is*, *are*, *was*, *were*, and *am*, but not *be*.[5] Although *been* and certain forms of *have* are also learnable by CAM, they won't be addressed further here since they are irrelevant to the subject of this paper.

Table 3 therefore summarizes the target orderings CAM should be able to learn.

Input

The relevant excerpt of a typical input to CAM is the following:

```
((the) (man)          (is) (walk ing))
(()    (3rd-person ()    (present
        singular)         progressive))
```

The first line shows the linguistic string, which is segmented into words. If a word contains inflectional morphemes, these are segmented from their stems.

The second line of the input shows the features of the perceptually salient words in the sentence. These features are assumed to be perceptually recoverable: a vision system capable of the object recognition required for this task could presumably recover these additional features with little difficulty.

The input file that generated Table 5 consists of a sequence of 24 inputs as just described. While this input format may seem a bit simplistic, it is sufficient for the acquisition of most of the subject-verb agreement rules in English. Also, CAM's inputs needn't be watered down in order to avoid the occurrence of a construction that might interfere with its ability to construct a grammar. CAM is capable of filtering out inputs too complex to handle. The ability to filter out irrelevant features, such as those encoding animacy or shape in English for example, has not yet been implemented in CAM, but there is no theoretical problem with using the Relevance Property [Nicholl and Wilkins, 1990] to do so, since no morpheme that falls under the scope of the Relevance Property is relevant for animacy or shape in English. The speed of the Relevance Algorithm used to implement the Relevance Property is such that many such irrelevant features can also be filtered out.

Procedure

CAM is actually a system of several learning procedures each devoted to learning a different aspect of language. Before agreement learning proceeds, categories are formed and CFG rules are built: this is why the whole CAM system is necessary; a Boolean learner cannot handle syntax independently. For example, the formation of one closed category (AUX) is discussed in

[5] This is simply because modals have not yet been given to CAM in experimental trials; CAM has no theoretical learnability problems with modals.

1. Increase the length of the conjuncts one term at a time.
2. Consider all hypotheses of the same length simultaneously.
3. Collect all single-term hypotheses involving noun features into a single conjunction prior to Boolean learning.

Table 4: Boolean learner constraints. These constraints are sufficient to enable a Boolean learner to duplicate Brown's complexity orderings. See text.

[Nicholl and Wilkins, 1990]. Agreement rules are constructed after an input like that above in section 3.2 is fully parsed by previously learned CFG rules.

Once an input can be unambiguously parsed by previously acquired CFG rules, the features in line 2 of the input (see section 3.2) are percolated (or projected) to the maximal category of the word they correspond to. CAM uses a traditional grammar formalism, so the features (3rd-person singular), which correspond to man, get propagated all the way to NP. A different grammatical theory might propagate to N^{max} instead.[6]

Once the features are propagated, they are read off the maximal categories in the sentences along with their syntactic positions and collected together into a set. Then, the grammatical morphemes already learned and recognized by CAM are collected into another set. So in the case of the example input in the previous section, the feature set would look like the following, although with more precision as to the exact syntactic position of the elements in question:

```
(NP 3rd-person)
(NP singular)
(V present)
(V progressive)
```

According to CAM constraint 3, the noun-related terms are collected together into one conjunction, shown below. This not only leads to faster acquisition, it also leads to more accurate results than otherwise, since if this is not done, a "race condition" can develop between noun-related terms and verb-related terms, occasionally leading to acquisition orders that violate Brown's orderings.

```
(and (NP 3rd-person) (NP singular))
(V present)
(V progressive)
```

[6] 'NP' is of course a more traditional representation for 'Noun Phrase', while 'N^{max}' is from X-bar theory. My intention here is to keep the exposition as general as possible, and avoid making a commitment to any specific syntactic theory so workers in different theories can still make use of what is presented here.

The set of grammatical morphemes would look like this:

```
verb-ending = -ing
AUX = is
```

The term verb-ending is just for exposition; in actuality an internal syntactic representation is used instead. Same for AUX. I shall avoid more precision in the interest of greatest generality.

Results

```
----------------------
(= (S 4 V 3 R 1) ING)
(= (S 2 V 2 R 1) ED)
(= (NP 3 N 2 R 1) S)
(= (S 4 AUX 2) (AM))
(= (S 2 V 2 R 1) S)
(= (S 4 AUX 2) (IS))
(= (S 4 AUX 2) (WAS))
(= (S 4 AUX 2) (ARE))
(= (S 4 AUX 2) (WERE))
----------------------

----------------------
(= (S 2 V 2 R 1) ED)
(= (S 4 V 3 R 1) ING)
(= (NP 3 N 2 R 1) S)
(= (S 4 AUX 2) (IS))
(= (S 4 AUX 2) (AM))
(= (S 2 V 2 R 1) S)
(= (S 4 AUX 2) (WAS))
(= (S 4 AUX 2) (ARE))
(= (S 4 AUX 2) (WERE))
----------------------
```

Table 5: Edited output of CAM showing the order in which it acquires morphemes.

One can see clearly that this algorithm will duplicate Brown's acquisition order formalism, since his results show that children learn shorter conjunctions before longer ones. For example,

$$\text{3rd-person reg.}(x+y) < \text{Uncont. Aux}(x+y+z)$$

is not only (1) one of Brown's acquisition order results obtained from observations of children (Table 3), but (2) the order resulting from the CAM constraints above (see Table 5). All the other results in Table 3 follow from the CAM constraints in the same way.

Now the actual runs of CAM shown in Table 5 manifest the CAM constraints as well as Brown's acquisition order results (Table 3). This is what we wanted to demonstrate.

Discussion and Future Work

This paper has shown how the three CAM constraints are sufficient to duplicate Brown's complexity orderings when applied to a Boolean learner embedded in a

larger language acquisition program. We suspect that prior work in formal language acquisition, e.g., [Anderson, 1983], [Pinker, 1984], cannot duplicate Brown's results because these systems do not obey CAM constraints 2 and 3. Both systems examine one feature at a time in sequence; if this sequence happens to be incorrect, their acquisition orders will deviate from Brown's results. CAM constraint 2 prevents this from happening.

```
verb-ending: ING
     (V PROGRESSIVE)

verb-ending: ED
     (V PAST)

verb-ending: S
    (((AND (NP 3RD-PERSON)
           (NP SINGULAR)
           (V PRESENT))

AUX: AM
    (AND (NP 1ST-PERSON)
         (NP SINGULAR)
         (V PRESENT))

AUX: WERE
    (((OR
        (AND (NP 3RD-PERSON)
             (NP PLURAL)
             (V PAST))
        (AND (NP 2ND-PERSON)
             (NP PLURAL)
             (V PAST))
        (AND (NP 2ND-PERSON)
             (NP SINGULAR)
             (V PAST))
        (AND (NP 1ST-PERSON)
             (NP PLURAL)
             (V PAST)))
```

Table 6: Some of the morpheme rules learned by CAM. Lightly edited.

The CAM constraints are not *logically necessary* constraints because there are potentially many other constraints (or algorithms) that could still duplicate Brown's results. One possibility would be to simply to replace CAM constraints 1 and 3 with: "add exactly one verb feature at a time." This might work, since in all of Brown's orderings (see Table 3), the only differences are due to the presence or absence of (1) the past feature and (2) the progressivity feature.

The most obvious direction for future work would be to extend these results to all 14 of Brown's morphemes. The irregulars are a clear target, in that it might be possible to extend the input to the Boolean learner to include lexical forms. Possessives might come next, especially if CAM's semantic input is extended: presently CAM has no way of representing possession or any other adjectival property. Finally, the copulas and prepositions might be mastered, but only after cross-linguistic research has identified the various syntactic realizations of states, and whether there might be tests for these.

References

[Anderson, 1983] John R. Anderson. *The Architecture of Cognition.* Cambridge, MA: Harvard, 1983.

[Berwick, 1985] Robert C. Berwick. *The Acquisition of Syntactic Knowledge.* Cambridge, MA: MIT Press, 1985.

[Brown, 1973] Roger Brown. *A First Language: The Early Stages.* Cambridge, MA: Harvard University Press, 1973.

[Chomsky, 1965] Noam Chomsky. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press, 1965.

[Gold, 1967] E. Gold. Language identification in the limit. *Information and Control*, 16:447–474, 1967.

[Lyons, 1977] John Lyons. *Semantics*, volume 2. Cambridge: Cambridge University Press, 1977.

[Nicholl and Wilkins, 1990] S. Nicholl and D. Wilkins. Efficient learning of language categories: The closed-category relevance principle and auxiliary verbs. In *The Twelfth Annual Conference of the Cognitive Science Society*, 1990.

[Pinker, 1984] Steven Pinker. *Language Learnability and Language Development.* Cambridge, MA: Harvard, 1984.