On the Possibility of *Lexical* Bootstrapping

Introduction

It has been observed many times that, in English and many other languages, verb learning starts later and proceeds more slowly than noun learning (e.g., Macnamara, 1972). It has been speculated that this is because verbs are generally harder to learn than nouns, at least in part because the word-world relations between verbs and their referents are harder to detect than those between nouns and their referents (e.g., Gentner, 1982). A great deal of research in language acquisition has therefore focused on the idea that children might use intra-linguistic cues to verb meaning as part of the verb learning process. In particular, it has been shown that children can use *syntax* to guide their acquisition of verbs – a strategy known as *syntactic bootstrapping* (Gleitman, 1990).

Syntactic bootstrapping, however, is limited. It requires that the child already have the ability to identify the syntactic roles in an utterance and associate them with semantic roles. Some approaches presuppose that knowledge of both semantic roles and linking rules (rules linking semantic roles to syntactic roles) is innate. That knowledge can then be used, by a procedure known as *semantic* bootstrapping (Pinker, 1984), to discover syntactic roles. Only then is the child in a position to use syntactic bootstrapping to help learn verb meanings.

In this work, we ask whether it is possible that there might be a more direct way of using intra-linguistic cues to assist in the acquisition of verb meanings. In particular, we consider whether there are informative co-occurrence relationships between nouns and verbs in parental speech to children. By *informative*, we mean that the statistical patterns of co-occurrences between nouns and verbs bear systematic relations to the meanings of those verbs. That is, the question we seek to answer is whether knowing which nouns a verb tends to co-occur with could give the child a clue as to the meaning of that verb. If so, then the child could use this information as an aid to verb learning. Because that strategy would only require being able to identify the lexical categories of words, not their syntactic or semantic roles, it could potentially be useful earlier in the verb learning process than syntactic bootstrapping.

In this work, we assess whether the language learning environment provides enough information for such a procedure to be successful.

Method

To determine whether there are informative noun-verb co-occurrences in parental speech to children, we performed a corpus analysis using the CHILDES database (MacWhinney, 2000), a large set of corpora of transcribed speech between parents and children. It has recently been released with a full set of morphosyntactic codings for all English corpora, allowing us to automatically detect nouns and verbs. We used all of the corpora in the American English portion of the database. There are a total of 36 separate corpora. The ages of the children studied in these corpora range from 6 months through 10 years, although the majority of them lie in the range from 1 to 5 years. There are 889 children in total.

Our analysis was conducted by a computer program that considered every transcript in every corpus from the CHILDES American English collection. For each transcript, we attempted to identify the "target child" (i.e., the child to whom parental utterances were directed). In transcripts where there was only one child, it was assumed that child was the target child. In transcripts where more than one child was represented, we used the participant ID header to identify the target child. Transcripts for which a target child could not be identified were excluded. We also attempted to identify parents in the transcript, using the roles "Mother" and "Father" in the participant ID header. Transcripts for which a parent could not be identified were excluded.

Next, we went through each utterance in the transcript. Utterances not produced by the target child or a parent were ignored. For each utterance included in the analysis, we examined the morphosyntactic coding in order to identify nouns and verbs. For each verb in the utterance, we recorded the closest noun to the left and the closest noun to the right. For example, the following utterance:

```
*MOT: I was gon (t)a ask you where
you got (th)em .
```

which has the MOR line:

```
%mor: pro|I v|be&PAST&13S v|go-
PROG inf|to v|ask pro|you
adv:wh|where pro|you v|get&PAST
pro|them .
```

would be counted as shown in Table 1:

	Before Verb	After Verb
pro|I	3	0
pro|you	1	3
pro|them	0	1

Table 1. An example count of nouns before and after verbs from a parental utterance in CHILDES.

This may be easier to interpret visually, as in Figure 1.

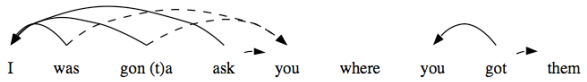

Figure 1: Nouns before and after the verbs in an example sentence from CHILDES. Lines are directed from the verbs to the corresponding nouns. Nouns before verbs are indicated by solid lines from right to left. Nouns after verbs are indicated by dashed lines from left to right.

Results

The most common nouns, by a long margin, in parental speech to children are pronouns. More interesting, for our purposes, is the fact that there *are* informative correspondences between pronouns and other nouns, on the one hand, and broad verb class, on the other.

Principal components analysis and clustering over the distribution of verb-noun co-occurrences reveal that there is a class of verbs that tend to occur most frequently before the pronoun *it*, and that there are classes of verbs that tend to occur most frequently after the pronouns *I* and *you*. Moreover, these "lexically defined" verb classes correspond to important broad semantic verb classes.

Roughly speaking, when a verb is frequently followed by the pronoun *it*, then the verb tends to have to do with physical motion or transfer. Verbs not frequently followed by *it* tend to be mental state verbs. Among these, those that frequently occur after the pronoun *I* mostly have to do with knowledge, whereas those that occur after the pronoun *you* mostly have to do with desire – an interesting consequence of the parent-child ecology.

Conclusions

It appears that, indeed, there *are* informative co-occurrence relationships between nouns and verbs in parental speech to children. That is, there are classes of lexical co-occurrences in parental speech to children that correspond to broad semantic verb classes. We suggest that children *could* use these co-occurrences to do "lexical bootstrapping" – a child knowing that most of the verbs that she has heard followed by *it* are physical verbs, may suppose that a novel verb followed by *it* is also a physical verb. Likewise, knowing that most of the verbs that she has heard preceded by *I* and *not* followed by *it* have been about knowledge, the child may suppose that a novel verb preceded by *I* and not followed by *it* is about knowledge. In other words, the present study has shown the information necessary for children to use lexical bootstrapping to discern the broad semantic class of unknown verbs is present in the input. Experiments are currently underway to test the hypothesis that children actually do use this information – that learning verbs used in frames like "He's blicking it" is easier than learning those used in frames like "John is blicking the box".

References

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Language, thought and culture* (Vol. 2, pp. 301-334). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition, 1*(1), 3-55.

Macnamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review, 79*, 1-13.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed. Vol. 2: The Database). Mahwah, NJ: Lawrence Erlbaum Associates.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.