The Role of Lexical Acquisition and Segmentation in the Acquisition of Phrase Structure

Helen Seville & P. Hancox

The argument advanced here is that the acquisition of syntax cannot be considered apart from those processes which determine the units upon which it operates during the course of acquisition, namely, lexical acquisition and segmentation. Peters (1983) has argued in favour of such an approach to acquisition, based upon an extensive analysis of segmentation in child language. Here, it is identified as the solution to a formal problem addressed in developing a computational model of child language acquisition. This is the problem of how phrase structure may be acquired.

Existing computational models have acquired phrase-structure grammars by assuming the innateness of syntactic categories and the X-Bar Theory of Phrase Structure (e.g., Berwick 1985). However, such models are too powerful to capture the observed stages in children's gradual acquisition of language. Models which have not assumed innate linguistic knowledge have acquired only finite-state grammars (e.g., Hill 1983), whereas the target grammar is a phrase-structure grammar (Chomsky 1957, pp.21-32). An account of phrase structure acquisition is, therefore, required.

The account offered is in the form of a re-statement of a problem which appears intractable, given the assumptions of existing models. These are the 'simplifying' assumptions that the learner is able to segment utterances into their constituent morphemes and has acquired the meanings of certain content words. A finite-state grammar is assumed as the initial state, e.g.,

\[
S \rightarrow \text{Det Noun Verb Prep Det Noun}
\]

Acquiring a phrase-structure grammar may be characterized as involving, as a first step, acquiring the noun phrase rule in which the constituents are lexical categories, i.e.,

\[
S \rightarrow \text{Np Verb Prep Np}
\]

\[
\text{Np} \rightarrow \text{Det Noun}
\]

In the absence of X-Bar Theory, the kind of information available to the learner is semantic. Recognising that the determiner and noun function as a single unit is, therefore, only possible given the assumption that the meanings of functional morphemes, like determiners, have been acquired. The difficulty is that phrase structure appears to be exactly the kind of local information needed to guide functional morpheme acquisition, assuming limited working memory.

In the re-statement of the problem, no lexical acquisition or segmentation are assumed. Inputs to learning consist of utterances paired with their semantic representations which, in the earliest stages of language learning, are acquired as lexical entries. Segmentation acts upon lexical entries which contain a common phonological string, resulting in the creation of new lexical entries corresponding to units like "the cat". These are utilized in the first rules acquired, so that in the initial finite-state grammar the 'noun phrase' exists as a lexical category, i.e.,
S  -->  Np ...
Np  -->  "the cat"

Further segmentation breaks down units like "the cat" into their constituent content words and functional morphemes. When grammar rule acquisition utilizes the new lexical entries acquired, a phrase-structure grammar emerges, i.e.,

Np  -->  Det Noun
Det  -->  "the"
Noun  -->  "cat"

The approach to acquisition outlined above has been implemented in a computational model, demonstrating that it is not necessary for models of child language acquisition to assume innate linguistic knowledge such as X-Bar Theory. One motivation for rejecting innatist assumptions is the aim of developing a model which captures some of the observed features of child language acquisition. The remainder is, therefore, devoted to a discussion of one aspect of learning in the model, functional morpheme acquisition.

A number of stages may be observed in children's acquisition of functional morphemes. Initially, there is a stage of largely correct but unproductive usage. This is followed by the systematic omission of functional morphemes. Eventually, rules for their use are acquired. Segmentation in the model developed suggests a novel account of the transition from the first to the second of these stages.

The meaning of a lexical entry is represented by a number of feature: value pairs, e.g.,

"the cat" [pred:cat, num:sg, def:pos]
"the dogs" [pred:dog, num:pl, def:pos]

Segmentation involves an inductive inference, the basis of which is a number of such lexical entries which share a common phonological string (e.g., "the") and feature: value pair (e.g., 'def:pos'). A new lexical entry is created corresponding to the common unit, i.e.,

"the" [def:pos]

The meaning of a unit such as "the cat" is the result of the unification of the meanings of its constituent units, "the" and "cat". It can be deduced from the nature of unification that different feature: value pairs in the lexical entries acted upon by segmentation derive from their different constituent units, i.e.,


However, it cannot be inferred that these are the only features in the lexical entries acquired corresponding to the different units. Shared feature: value pairs may derive from these as well as from the common unit (as is the case in agreement), i.e.,

[def:pos] U [] = [def:pos]
[def:pos] U [def:pos] = [def:pos]

In the model it is assumed that, at the point of segmentation, the lexical entries of the different units retain all the features of the original lexical entries, i.e.,

"dogs" [pred:dog, num:pl, def:pos]
"cat" [pred:cat, num:sg, def:pos]
The basis for this assumption is another aspect of learning in the model whereby the meanings of lexical items are acquired through the removal of features discovered to be redundant, rather than through the addition of features. The suggested explanation of functional morpheme omission is that, until redundant feature:value pairs are removed from the lexical entries of content words, functional morphemes appear to be redundant.

REFERENCES