**Methods for Assessing Children's Syntax**

edited by
Dana McDaniel,
Cecile McKee, and
Helen Smith Cairns

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by Asco Trade Typesetting Ltd., Hong Kong.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Methods for assessing children's syntax / edited by Dana McDaniel, Cecile McKee and Helen Smith Cairns.
p. cm. — (Language, speech, and communication)
Includes bibliographical references and index.
ISBN 0-262-13325-3 (hc : alk. paper)
1. Language acquisition—Research—Methodology. 2. Grammar, Comparative and general—Syntax—Research—Methodology.
I. McDaniel, Dana. II. McKee, Cecile. III. Cairns, Helen Smith.
IV. Series.
P118.15.M48 1996
401'.93—dc20 96-33915
CIP

# Chapter 1

# Collecting Spontaneous Production Data

Katherine Demuth

## 1.1 Introduction

Much of the earliest work on child language acquisition took the form of longitudinal diary studies, where parents documented developments in their child's grammar and/or lexicon (e.g., Stern and Stern 1907; Grégoire 1937, 1947). Later, with the emergence of tape-recording technology, both parents and nonparent researchers were able to collect spontaneous speech samples from a variety of children. This paved the way for a significant increase in both the amount of material that could be collected and the types of research issues that could be addressed. Many of these issues, such as the path to development of grammatical competence, the contributions of general cognitive abilities, and the role of input, continue to be hotly debated today, not only by linguists and researchers working on language acquisition, but also by learning theorists and cognitive scientists more generally.

Along with a growing interest in the nature of linguistic structure (Chomsky 1957, 1965) came an increasing concern with how these structures are actually acquired. Some of the earliest research on the acquisition of English used spontaneous production data to begin to address this question (e.g., Braine 1963; Brown and Fraser 1963; Miller and Ervin 1964; Bloom 1970). It was also recognized that crosslinguistic data are essential for understanding the nature of language acquisition. This led Slobin and colleagues to the development of *A Field Manual for Cross-Cultural Study of the Acquisition of Communicative Competence* (Slobin 1967). Several studies of children learning other languages followed (Finnish, Bowerman 1973; Samoan, Kernan 1969; and Japanese, McNeill 1966a, McNeill and McNeill 1966). Since that time, the collection of spontaneous production data has become a frequently used method for

addressing acquisition questions, and the number of crosslinguistic studies using this technique has continued to grow (e.g., Slobin 1985b, 1992). Many spontaneous production corpora from a variety of languages have been computerized, and an increasing number are available as part of the CHILDES data archive at Carnegie Mellon University (MacWhinney and Snow 1985; MacWhinney 1991). The collection of spontaneous data has already made a significant contribution to language acquisition research. It is not, however, to be undertaken lightly: spontaneous production data are useful only when collected systematically and with careful attention to details that affect the quality of the resulting corpus.

One set of spontaneous production data that has had a significant and continuing impact on the field has been Roger Brown's longitudinal study of the English-speaking children given the pseudonyms Adam, Eve, and Sarah (Brown 1973). This data set continues to be useful because it was carefully collected and documented, because it provides longitudinal evidence for similar stages of development across three children with different developmental rates, and because data collection took place during the morphosyntactically interesting period when the mean length of utterance (MLU) was between 1.75 and 4 morphemes. Although the specific goal of Brown's study was to examine English-speaking children's development of grammatical morphology, these corpora continue to provide researchers with a rich set of production data that can be used to investigate many syntactic issues. For example, they have been used by Stromswold (1990b) in investigating children's acquisition of auxiliary verbs, by Marcus et al. (1992) in examining morphological overgeneralization, and by Bloom (1990) in a treatment of children's subjectless sentences. When collected appropriately, spontaneous production data can provide a wealth of information to be tapped repeatedly over the years. In the following section I discuss the kinds of syntactic phenomena that can most profitably be examined using this type of data.

### 1.2 Syntactic Phenomena Investigated

A primary goal of language acquisition research has been to assess the Chomskyan notion of grammatical competence. It is often more difficult to assess young children's knowledge of language than adults'. Researchers have therefore devised various methods appropriate for assessing young children's early grammatical abilities, and many of these

are discussed in later chapters (see chapter 11, this volume). Spontaneous production data can also be used to determine certain types of grammatical competence, especially in the area of morphosyntactic development.

### 1.2.1 Pro-Drop and Parameter Setting

Since the early and mid 1980s grammatical morphology has played an increasingly important role in the construction of syntactic theory. This state of affairs has been reflected in the questions researchers have asked about the course of language acquisition. For example, in the development of the Principles and Parameters approach to linguistic structure (Chomsky 1981), it was noted that in some languages (e.g., English) an overt subject is obligatory, whereas in others (e.g., Italian) it is not. Hyams (1986) suggested that the lack of pronominal subjects in early English was evidence of a null-subject stage of development, where young English speakers' initial setting of the pro-drop parameter was hypothesized to be similar to that of null-subject Italian. Spontaneous speech data from English-speaking children have subsequently been used to argue against this view (e.g., Valian 1991) by providing statistics on how frequently young English speakers use lexical and pronominal subjects.

### 1.2.2 Functional Categories and Syntactic Structure

Grammatical morphology and its role in children's developing grammars have taken on renewed relevance as the distinction between functional and lexical categories (closed- vs. open-class items) has moved into the mainstream of syntactic theory (Abney 1987; Chomsky 1991). A flurry of research activity has ensued examining spontaneous production data from languages as diverse as Italian, English, Swedish, German, Swiss-German, French, Korean, and Sesotho (see Meisel 1992; Lust, Suñer, and Whitman 1994; Hoekstra and Schwartz 1994; and references therein). Researchers have studied how and when children acquire various aspects of grammatical morphology, including the marking of tense, person, number, gender, and case, as well as the placement and use of auxiliaries, negation, determiners, and complementizers. Some of these studies have drawn on original findings from Brown's corpora: Bellugi (1967) studied the emergence of children's use of negation and subject-auxiliary inversion, and Brown (1968) investigated stages in the acquistion of yes/no questions and *wh*-questions. More recently, spontaneous production data have been used by Pierce (1992) and Déprez and Pierce (1993) to investigate

negation in French and by Radford (1994) to examine the syntax of early English *wh*-questions. I have also used spontaneous production data from Sesotho (a Bantu language) to explore the development of complementizers and the formation of relative clauses, questions, infinitival complements, and embedded clauses (Demuth 1995).

### 1.2.3 Passives, Causatives, and Grammatical Relations

Spontaneous production data have also been used to explore how and when passives, causatives, and other grammatical-function-changing operations are acquired. Although passives rarely occur in the spontaneous speech of English-speaking children, they appear much more commonly in the spontaneous speech of children learning Bantu languages (Sesotho, Demuth 1989, 1990; Zulu, Suzman 1985). Children also use ergative marking and antipassive constructions quite early in learning languages such as K'iche' (Pye 1992) and Inuktitut (Allen and Crago 1993). Such findings have called into question previous theoretical notions of grammatical complexity and children's early grammatical abilities. Other studies, including work on the acquisition of causative constructions (cf. Bowerman 1982), shed light on the child's developing lexicon and on lexical interactions with syntactic development. Much of this latter research draws on longitudinal diary studies of children's spontaneous productions and focuses on overgeneralization errors.

### 1.2.4 Morphological Paradigms and Learning

Spontaneous production data such as Brown's (1973) corpora have also been used in addressing learnability issues such as how seemingly complex inflectional paradigms are learned (e.g., Rumelhart and McClelland 1986; Pinker and Prince 1988). Issues of input become extremely important in such studies, and researchers are beginning to reexamine spontaneous production corpora, looking more closely at the distributional properties of the input and its relationship to the acquisition of morphological paradigms (e.g., Clahsen et al. 1992; Ziesler and Demuth 1995).

In sum, the use of spontaneous production data has been and continues to be extremely important for addressing various issues relating to morphological and syntactic development. As technological and theoretical advances in the area of "corpus-based" linguistics increase, so will the advantages of using spontaneous production data to address acquisition and learnability issues.

## 1.3 Spontaneous Production Data Collection Procedures

Like any other type of data collection, spontaneous production data collection is useful only if collection methods are carefully planned. Planning must include consideration of both the research questions to be asked and the methods to be used in the process of data collection itself. Given the labor-intensive nature of collecting and coding spontaneous production data, it is advantageous to have both short-term and long-term research goals in mind. This should hold not only for the specific research topic(s) to be addressed, but also for issues relating to the number of children, the ages of the children, the length of the study, the frequency and length of the recordings, and the conditions of the recording situation, including the site, interlocutors, and acoustic quality of the recording itself. Each of these issues is discussed in more detail below.

### 1.3.1 Number of Children to Include in a Study

Acquisition studies have shown that the course of language development varies to a certain extent from child to child. Although much of this variation is related to when certain constructions are acquired rather than to the course of acquisition, it is generally accepted that a study of several children is more informative than a study of one child. It is therefore preferable to collect spontaneous production data from more than one child. Given a target of three children, it may be advisable to start a study with four. This is especially important in research settings where children and their families may move away before the completion of a longitudinal study, or succumb to sickness or death, as may happen in communities with high early childhood mortality. Furthermore, one or more children or families may drop out of the study for reasons of work, frustration, or other priorities. Brown's (1973) study of three children provides a nice sample of variation, where Eve is much more precocious than either Adam or Sarah. Such diversity is vital to constructing a coherent theory of acquisition.

### 1.3.2 Age Range of the Children and Longitudinal Scope of a Study

The age range of the children to be recorded and the length of the study should be determined on the basis of the general research questions and the specific grammatical phenomena being investigated. Given individual variation in development, a certain amount of variation in the age of the

children studied should be allowed: individual children's MLU may be a more accurate measure of linguistic ability than age (Brown 1973). This is true even though there may be difficulty calculating MLU crosslinguistically, especially for highly inflected languages such as Hebrew (Dromi and Berman 1982) and West Greenlandic Fortescue 1985; Fortescue and Lennert Olsen 1992).

If little previous acquisition work has been done on the language under study, it might be advisable for the researcher to consult persons in the community who are knowledgeable about child language, or to listen to children of different ages to determine if certain constructions are in use. In general, however, children between the ages of 2 and 3 show rapid phonological, lexical, morphological, and syntactic development. If the study concerns development of grammatical morphology, it is advisable to begin recordings with children younger than 2 years in order to catch the transition stage. If grammatical constructions such as passives, relative clauses, or complementation are to be examined, the study should include older children, perhaps between 2;6 and 4 years. If the study looks at certain types of lexical categorization involving complementation and argument structure relations, children between the ages of 3 and 5 should probably be included.

In situations where it is impossible to follow one set of children for longer than 12 months, it may be useful to collect data from children in one or two age groups, or from children of overlapping ages (e.g., 3;6–4;6, 4–5, 4;6–5;6 years). This may be especially useful when initiating the study of a language where little or no previous acquisition work exists and it is unclear when children acquire certain constructions.

### 1.3.3 Selecting Children for a Study

Several factors should be considered in selecting children to participate in a longitudinal study of spontaneous speech production. First, if the community is bilingual or multilingual, the language situation in the home and/or day care center should be carefully assessed to ensure that the monolingual/bilingual setting is appropriate to the requirements of the study. This may be a determining factor in selecting the initial research site. For my work in Lesotho, in southern Africa, I decided to base my study in a rural village rather than an urban center to avoid possible English influence on the children's acquisition of Sesotho (Demuth 1984, 1992). Second, it is good to have a gender balance among the children in the study, so that sex-based rates of maturation and gender-based use of

language (in some cultures) can be represented. Third, children with a history of ear infection and/or other health problems and children with obvious cognitive deficits should not be included unless the research is specifically designed to study language development in these populations; both cognitive deficits and health problems that affect hearing may have a significant negative impact on children's language development.

Once the age range of the children to be studied has been determined, the researcher should visit several children in the community to determine which children and families are most appropriate for inclusion in the study. These visits are useful in two respects. First, they offer the researcher an opportunity to become familiar with some children and their families. Second, they provide a basis for deciding which children will become part of the study. If MLU is a factor in selecting children for the study, this period of familiarization can facilitate assessment of children's stage of linguistic development. Finally, the families as well as the children will be involved in the study: the researcher will have to arrange times convenient for recording, and if parent-child interactions are required, the parents will have to agree to participating in the research themselves. In some research situations, such as with the Inuit in Canada (Crago 1988), parents work during the day, and recordings have to be carried out with the cooperation of other caregivers. Prerecording visits to families therefore provide the researcher with critical information regarding which families and children will be most appropriate for the study. It is important that the researcher feel at ease with both the families and the children; the quality of the data will be adversely affected if recording sessions are stilted or artificially constructed in any way (see Clark 1982).

### 1.3.4 Frequency and Duration of Recording Sessions

A decision must be made about the frequency and duration of recording sessions. In Brown's (1973) study, Eve was recorded for at least half an hour every week, and Adam and Sarah were recorded for about an hour every two weeks. In addition, more data were collected more frequently when morphosyntactic changes were occurring at a rapid pace. By contrast, for my work on Sesotho acquisition I collected data less frequently (once a month), but in a variety of discourse situations, resulting in much larger samples per session (3–4 hours). It is useful to have a plan for how often and how long to record, but it is also necessary to be flexible and ready to adapt when recording opportunities arise. It may be advisable to collect more data than actually needed to ensure that at least a certain

number of relevant utterances (i.e., utterances containing constructions of a certain grammatical type) are included in every recording session.

The collection of spontaneous production data is at best a sampling technique. An important consideration in determining how much material to collect is to ensure that the data constitute a "representative sample" of the child's productive language capabilities at the time. What counts as "representative" will depend greatly on the grammatical phenomena being studied and how frequently these constructions occur in everyday discourse. For example, more data are needed to examine complex grammatical constructions such as passives, relative clauses, and complementation; fewer data are needed to examine the use of subject agreement or other frequently occurring morphosyntactic phenomena. As will be discussed in the following sections, the recording site and recording procedures often have as much to do with collecting representative samples as do the frequency and duration of the recordings themselves.

### 1.3.5 The Recording Situation

Several factors, including the site of the recording sessions, the participants, the interactive situations being recorded, and the type of recording equipment used, all play an important role in the quality of the spontaneous production data collected. Many of these issues are similar to those of collecting experimental production data, though others are necessarily different. Each of these issues is discussed more fully below.

Most longitudinal spontaneous production studies take place in and around children's homes rather than in an acoustically treated laboratory. There are several reasons for this. First, the phenomena investigated using spontaneous production data have generally been of a morphological, syntactic, or semantic nature rather than phonological or acoustic. Second, it is generally recognized that young children are more likely to talk freely, and to use more grammatically complex linguistic constructions, when they are in a familiar environment. It is for this reason that studies using spontaneous production data, which have frequently involved upper-middle-class children, have focused on mother-child interaction as being the prototypically "familiar" setting in which the upper end of children's linguistic abilities would be readily observable. However, studies of children learning other languages in other cultural settings have found that children typically interact with a large range of both adults and children on an everyday basis and that recording should not necessarily be confined either to mother-child interactions or to one setting. For instance, in

rural Lesotho I found that grandmothers, peers, and older siblings were some of the most frequent interlocutors with young children and that mother-child interactions decreased significantly around the age of 2;6 years, with or without the birth of a younger sibling. In addition, some of the children's most advanced linguistic forms, such as restrictive relative clauses, occurred during peer and sibling interactions where children had to be extremely linguistically sophisticated to get what they wanted (Demuth 1984). Thus, although the home environment may be the site in which children feel most comfortable, that environment may include many more discourse participants than simply the mother. This may be especially true when extended families or peers live nearby, or when the child has older siblings. Such interactions can provide an extremely rich set of production data, from both the child and other caregivers, including fathers, aunts and uncles, grandparents, older siblings, and cousins. One of the challenges for the researcher is to determine, given a particular culture and specific family situations within that culture, which interactive situations are the most productive for collecting children's speech.

Interactions that involve either one or a number of participants may not necessarily be confined to one site. Some of the richest interactive and linguistic situations may be embedded in a range of daily activities including bathing, cooking, eating, and playing outdoors. Noise factors, such as water running into a bathtub, the TV, washing machine, or dishwasher in the background, rain pelting on a tin roof, cooking noises, loud music from next door, or ten preschoolers at a birthday party, can obliterate the speech of the target child; in such situations it is best to stop recording and continue later or the next day. The researcher should be flexible enough to take advantage of different recording opportunities as they arise. Allen (1994) reports that one of her richest recording sessions with Inuit children took place five hours away from home at the family's summer camp.

The picture that begins to emerge here shows the researcher gradually becoming "part of the extended family." Both researcher and family have to make decisions about how this relationship will be negotiated, and it is highly relevant to the quality of data collected. By living and working in a small village of 550 people in Lesotho, I was able to establish a relationship of daily interaction with three families, becoming a member of the extended community and someone the children saw and talked with frequently. My transitions into and out of families' homes, with or without the tape recorder, became normal events in the life of each child, allowing me to record whenever and wherever the collection of spontaneous