

**Methods for Assessing
Children's Syntax**

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

The MIT Press
Cambridge, Massachusetts
London, England



© 1996 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

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 acquisition 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

Methc

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

linguistic productions looked promising. Sometimes this meant joining families for a meal; other times it meant racing after children as they chased chickens or played tag. The resulting data set is grammatically extremely rich, providing an excellent assessment of children's syntactic abilities in a broad range of discourse situations.

The researcher will have to decide whether to be a participant in verbal interactions, or simply an observer. If the researcher is not a native speaker of the language under investigation, interactions should probably be limited. When recording in Lesotho, I rarely initiated conversation with the children, only answering when spoken to, or warning children against activities that might lead to bodily harm, such as falling into the fire pit or playing with a sharp knife. Bowerman (1973), in her study of Finnish children, also took this approach. Even when the researcher is a native speaker of the language being studied, the goals of the research may influence decisions about researcher participation. For example, if one of the goals is to examine the type of input adults provide to children, the researcher will want to keep interaction to a minimum. On the other hand, if the researcher wants to do some informal elicitation to probe for children's knowledge of certain syntactic constructions, specific types of interaction could play a useful role.

In addition to audio recording, the researcher should arrange to take contextual notes or video recordings. This is important because the details of the setting and activities of the participants are often essential to interpreting children's utterances. For example, the use of a relative clause in English may or may not be restrictive, and it may only be notes such as "child looks at three dolls, then picks up the tall one," or the equivalent observed on videotape, that can provide the information needed to evaluate children's use of such constructions. Written or videotaped contextual notes should be keyed to the counter on the tape recorder for easy and accurate retrieval.

1.3.6 Recording Equipment

Spontaneous production data may be collected in a variety of indoor and outdoor settings. The type of recording equipment used should be selected accordingly. Consultation with someone knowledgeable about professional recording equipment, as well as the conditions under which the researcher will be recording, is highly recommended. In general, however, recording equipment should be of high quality, but also portable. Fortunately, researchers today can choose from a number of light, portable

audio and video recorders of professional quality. As recording technology becomes increasingly sophisticated, decisions will have to be made about when to move to the latest technology. Again, the goals of the research become important in making this decision. If the project is a preliminary, exploratory study where the long-term preservation of tapes is not critical, traditional analog recordings are probably adequate. If, however, the study is large and potentially of significant long-term interest to the field, like Brown's (1973) study, the use of digital recordings, with long-term archival shelf life, should be seriously considered. In either case, high-quality tape should be used.

New equipment should *always* be tested before the researcher disappears to the field. This is especially true if it incorporates new technologies, since the researcher may need to become familiar with the equipment, and since new products may lack effective quality control. As an additional precaution, backup recording equipment should always be taken to the research site; recording equipment, especially under intensive use, has been known to break and/or suffer the consequences of wear and tear, especially when used outdoors.

More important than the choice of tape recorder is the choice of microphone. Again, knowledge about the recording site will play a significant role in determining which type of microphone is best. If the recordings are to take place primarily in one room, it may make sense to hang an omnidirectional microphone in the middle of the room so that all speakers' voices can be heard. However, if recordings will take place on the go, it is advisable to use either a handheld directional microphone, or a wireless broadcast microphone built into a vest that the child wears, or both if discourse interactions with other speakers are also desired. The quality of the microphones should be good. If recording is done outdoors, microphones should be used with a wind screen.

Given the nature of collecting spontaneous production data from young children, it may be best to power the recorder with batteries rather than relying on an electrical outlet. It is therefore wise to carry extra batteries—nothing is worse than being ready to record only to find that the batteries are dead! Likewise, carrying an extra tape is a good idea: if the child is having a particularly verbose day, it may be worth collecting more than the usual amount of data. Finally, at the end of each recording session it is advisable to verify that recording actually took place and that the tape is intelligible. All tapes should be marked with the recording date and with the age and name of the child.

1.4 Transcribing and Tagging Spontaneous Production Data

Several recent publications have described widely used conventions for transcribing and tagging spontaneous speech data and have discussed which procedures are most appropriate for different types of production data (e.g., MacWhinney 1991; Edwards and Lampert 1993). The purpose of this section is not to repeat the material found there but to offer a procedural perspective on these issues, with specific reference to the types of decisions that will need to be made.

1.4.1 Getting Ready to Transcribe

Once a recording session has been completed, the researcher should begin transcription as soon as possible. This ensures the maximum transmission of contextual information and transcript accuracy. It is probably best to make a copy of the original tape and to transcribe from the copy, keeping the original for archival use or for backup; transcription involves lots of going back and forth, and tapes sometimes break under the strain. If the original recording was digital, the copy can be analog, since most transcribing machines still use analog tape.

If possible, audio recordings should be transcribed directly into a computerized database so that the corpus can be easily used for analysis. The format used for transcription will again depend partly on the goals of the research. Some researchers may decide to transcribe data into a format compatible with the files in the CHILDES data bank at Carnegie Mellon University. CHILDES offers child-language-oriented search-and-analysis programs (such as CLAN) that can calculate MLU and collect statistics on the frequency of occurrence of certain constructions (see chapter 2, this volume, for further details). Furthermore, these search programs are available for both PC and Macintosh computers, complete with documentation (MacWhinney 1991). There may be cases, however, where researchers wish to customize transcription and coding into a format that is more readily usable for immediate research purposes. In this case it is still advisable to code data into some sort of database. Excel and 4th Dimension are databases that some researchers have found useful. Customized search programs can be written for these databases, and the data can be converted to the CHILDES format at a later date.

Each file, or transcript, should include information about the child and the recording situation, such as the child's name (using a preselected pseudonym), the child's age, the date and site of the recording, and partic-

ipants in the recording session (e.g., mother, siblings, other relatives, friends). If the researcher is not a native speaker of the language being investigated, it may be advisable to transcribe the tape in conjunction with the mother of the child or some other native-speaker adult who knows the child well and might have been present during some of the recording sessions.

Finally, the researcher will have to decide whether to transcribe only the child's utterances or to include both the child and other interlocutors. In my Sesotho corpus I have transcribed all speech from adults, peers, and others who were interacting with the target child. This interaction is invaluable for understanding the context of the discourse and, in some cases, for determining what the child was trying to say.

1.4.2 The Transcription Process

As Ochs (1979) so aptly notes, the very process of transcription has theoretical consequences. At every stage of the data collection and transcription process, certain details are lost, resulting in a product that preserves only certain types of information. Given that researchers generally use only the final transcript when conducting syntactic analysis (and when they use the CHILDES data bank, this transcript is often the only data source available), the type and quality of the information included in the transcript will undoubtedly bias our understanding of how language is acquired. Decisions about what to transcribe and how to transcribe it therefore play a critical role in the types of syntactic and related research questions the data can be used to address. Some of these issues involve the level of phonetic detail transcribed, the inclusion of relevant contextual information, and decisions about what constitutes an "utterance." These and related issues are examined in more detail below.

Given the nonlaboratory nature of most spontaneous production recordings and a research focus on lexical, morphological, and syntactic issues, a broad phonemic (rather than narrow phonetic) transcription is probably adequate. In many cases broad phonemic transcriptions have used the orthographic conventions of the language (e.g., Brown 1973; Bowerman 1973). However, a decision will have to be made about how to transcribe children's phonetically altered forms, and a description of these conventions should accompany the transcripts. Any phonetic information relevant to the syntax should be marked. For example, in languages that use lexical and/or grammatical tone, such as many Southeast Asian and African languages, tone may need to be marked to capture lexical,

morphological, or syntactic information. The presence or absence of grammatical function items, including vowel or consonant quality, can also be extremely relevant for addressing certain syntactic questions. Even when transcribing English it may be advisable to use some convention for encoding intonational contours that can capture contrastive stress. Thus, a broad phonemic transcription may need to be carried out with attention to some phonetic detail. In some cases this may involve the use of diacritics not readily available on a keyboard. The CHILDES manual (MacWhinney 1991) has a series of conventions for entering such cases (PHONASCII), and it may be appropriate to use them unless the researcher finds them inadequate in some respect.

Another decision to be made is how to break up conversation into "utterance"-level units for transcription purposes. Again, the choice of transcription technique will vary depending on the research questions being asked. Much of my early work on Sesotho dealt with passive constructions; I therefore coded data according to clauses. Since the material on Sesotho exists in database format, it is possible to recover the complete utterances so that relative clauses and embedded complements can also be examined. Taking a different approach, Allen (1994) transcribed each utterance on a separate line and then counted the number of verbal clauses per utterance. A related issue is how to deal with "repetitions." In the Sesotho corpus I generally coded identical consecutive repetitions as one utterance, with a note in the "comments" column (a separate "field" in the database) about how many times it occurred. If "repetitions" were segmentally or prosodically different, or if utterances by other interlocutors intervened, I counted them as separate utterances. Transcript entries should all be keyed to the original tapes to facilitate easy access; the researcher may find it useful or even essential to return to the original tapes from time to time, either to check the original transcription or to transcribe additional information (such as phonetic or prosodic detail).

Contextual information often provides evidence of the pragmatic intent of the utterance, and this may influence the grammaticality of what was said. This becomes highly relevant to the syntactic investigation of focus constructions such as topicalization, clefting, relativization, the use of stressed pronouns, word order, and the like. Contextual information that has been captured in notes or on videotape should therefore be entered into a "context" field in the transcript.

Even with contextual information, a videotape, and the aid of a native speaker, it may occasionally be difficult to determine what the child actually said. Sometimes little or nothing is recoverable; in this case the researcher should indicate that the child said something, but that it was unintelligible. This will aid in understanding the nature of the discourse within an ongoing conversation. In other cases the child's utterance may sound like actual words, but it may not be clear if what the researcher hears is what the child actually said. In this case the researcher should note in a "comments" field of the transcript that there is uncertainty about what has been transcribed, and include alternatives if there are any. Some of these utterances may become disambiguated once more of the session has been transcribed. If not, and if a second transcriber is unable to shed light on the issue, the researcher may choose to disregard these entries or gloss them as unintelligible utterances.

Once a transcription has been completed, it should be checked and verified by another researcher. It may be advantageous if this person is a native speaker of the language, but one who was not present during the recordings. Verification should be conducted by listening to representative samples of the tape (perhaps 10% of the total data set) and retranscribing it. The two transcriptions should then be checked for validity. Backup copies of all work should be made.

Transcription is a painstaking process. I found that even when I worked with the mother or grandmother of the child, broad phonemic transcription of 1 hour of audiotape of the Sesotho corpus generally took 7 hours. Allen (1994) and Crago (1988) found that Inuktitut speakers transcribed about 2 to 5 minutes of videotape per hour. In other words, transcribing either audio- or videotape requires a large investment of time. The researcher should plan accordingly.

1.4.3 Tagging (Coding) the Corpus

Spontaneous production corpora are most useful if some type of tagging (grammatical coding) is included in the transcript (database). Again, the extent and type of tagging will depend, in part, on both the immediate and long-term goals of the project. Many corpora, such as the transcripts of Adam, Eve, and Sarah, have not been tagged. I have found that tagging is extremely helpful even in corpora from languages one knows well. For example, if one wanted to study the use of auxiliaries in an untagged English corpus, one would have to list all auxiliaries and then exclude

main verb uses of *be* and *have*. On the other hand, if auxiliaries have been tagged, a search for AUX will pull out all auxiliaries. Tagging becomes even more important when working with a lesser-known language—especially if the eventual goal is to make the corpus available to a larger audience, for example, by donating it to the CHILDES data bank.

In working with the Sesotho corpus, I have found it most fruitful to have separate fields for the child's utterance, the grammatical adult target form, a detailed set of morphological tags, and an English running gloss. The example in (1) gives an idea of how this can be done, where the different fields are Speaker = the speaker, Session = the recording session, Key = the counter number on the tape, Utterance = the child's utterance, Target = the adult equivalent target (i.e., what the child was trying to say), Tag = the morpheme-by-morpheme tag (grammatical coding) of the target utterance, English = a running English gloss that captures the meaning of the utterance, Context = contextual information, Comments = notable aspects of the utterance.

(1) Speaker	H
Session	IIA
Key	642
Utterance	ko rata
Target	ke-a-o-rat-a
Tag	1sSM-PRES-2sOM-like-IN
English	I like you
Context	child looks at doll
Comments	× 2

The transcript should then also include a separate glossary of all the tagging terms used throughout the corpus. For example, the glossary for the tags used in example (1) would include the items in (2).

(2) 1sSM	1st person singular subject agreement
2sOM	2nd person singular object agreement
PRES	present tense
IN	indicative mood
× 2	identical consecutive repetition of an utterance

The specific tags used will depend partly on the language being investigated and partly on the research questions being asked. This type of detailed tagging is extremely useful for conducting automatic searches of certain grammatical phenomena, especially in cases where children's pro-

nunciation of “words” or “morphemes” differs from the adult equivalent forms, or in cases of homophony. It is also useful in cases where the orthography of a language is not completely standardized and different transcribers use slightly different orthographic conventions. Furthermore, the inclusion of a field for “adult equivalent forms” provides other non-native-speaker researchers with ready access to where and how the child’s utterance deviates from the adult form. This information is often lacking in transcripts and in publications, making it difficult for both researchers and readers who do not have full command of the particular language to understand what the child has omitted or changed.

1.5 Disadvantages of Collecting and Using Spontaneous Production Data

Some potential problems involved with collecting spontaneous production data were discussed in section 1.4. Once the data have been collected, there are also certain limitations on what they can tell us about the course of acquisition. One of the central concerns in the field of language acquisition is to determine the nature of children’s underlying grammatical competence. Using production data to determine grammatical competence introduces certain problems of interpretation: how and when does the researcher know that the child has *productive grammatical competence* with certain grammatical forms? These issues are discussed below.

One limitation of using spontaneous production data lies in the nature of the sampling technique itself: if a particular grammatical construction does not occur in the sessions sampled, it is often difficult to determine the cause of its absence. For example, passive constructions occur relatively infrequently in young children’s spontaneous use of English. It was initially assumed that their absence was due to grammatical complexity or lack of linguistic “maturation” (e.g., Brown and Hanlon 1970; Borer and Wexler 1987). However, crosslinguistic evidence of early passives in spontaneous production data from languages like Sesotho indicates that English-speaking children should, in principle, be able to comprehend and produce passives by the age of 3 (Demuth 1989, 1990). Thus, spontaneous production data can provide positive evidence for the presence of a grammatical construction, but they are of limited use (without crosslinguistic evidence) in determining whether the absence of a particular grammatical construction is due to lack of linguistic ability, lack of exposure to the construction, or lack of appropriate discourse contexts in the sample.

It has long been realized that children's *comprehension* of some grammatical constructions, especially those pertaining to grammatical morphology, may precede children's *production* of these forms (e.g., Shipley, Smith, and Gleitman 1969). This is especially relevant to the current debate concerning the presence or absence of functional categories and the projection of syntactic structure (see Meisel 1992; Lust, Suñer, and Whitman 1994; Hoekstra and Schwartz 1994). Researchers using spontaneous production data may actually underestimate children's grammatical competence, especially at early stages of development. Spontaneous production data can often provide evidence of children's competence with certain constructions, but finding this evidence may require careful investigation on the part of the researcher. For example, early evidence of person marking in Sesotho comes from tonal evidence rather than from the presence of agreement morphemes, which tend to be phonologically reduced (Demuth 1993).

On the other hand, if a specific construction or grammatical item is present in spontaneous production data, it may be difficult to determine if its occurrence is "productive." For example, some researchers have argued that children initially have limited control of relative clauses (e.g., de Villiers et al. 1979; Tavakolian 1981a) and long-distance *wh*-movement (de Villiers, Roeper, and Vainikka 1990). Furthermore, some grammatical morphemes may initially be produced as lexicalized rather than productive forms. The researcher must therefore look for signs of "productivity," including morphological "errors" such as the overgeneralization of past tense *-ed* (e.g., *goed*, *catched*) in English. Experimental techniques (like those discussed in this volume) can often provide a more detailed assessment of children's linguistic competence with grammatical morphology and syntactic/semantic phenomena such as anaphoric relations, quantifier scope, island constraints, and the use of embedded and control structures.

1.6 Advantages of Collecting and Using Spontaneous Production Data

The greatest advantage of using spontaneous production data is that they can supply a wealth of information about many aspects of children's grammatical development. Longitudinal spontaneous production studies are particularly useful in identifying general developmental trends, providing an excellent picture of the overall course of development for a given language. This is especially helpful when initiating the study of a language on which little or no previous acquisition research has been

done. For example, passive constructions, which were initially thought to be difficult to acquire, turn out to be productively used in the spontaneous speech of 3-year-old speakers of Bantu languages like Sesotho (Demuth 1989, 1990) and Zulu (Suzman 1985). Furthermore, spontaneous production data from Inuktitut (Allen 1994) and K'iche' (Pye and Quixtan Poz 1988) show early acquisition of antipassive constructions in these ergative languages. Spontaneous production data can be used to assess children's grammatical competence in a number of ways: evidence of "productivity" comes from spontaneous overgeneralizations (e.g., regular past tense and plural marking on irregular English verbs and nouns), children's use of other novel forms that they could not have heard, the use of alternating forms (e.g., verbs with various endings), and children's own self-corrections (see Demuth 1989 and Allen 1994 for discussion).

Spontaneous production data that include utterances from interlocutors are especially useful in providing information about how frequently specific grammatical constructions typically occur in a language. They can therefore provide important evidence for determining whether a particular grammatical phenomenon, such as the passive, is linguistically difficult for young children to acquire or whether it simply fails to appear because of language-particular discourse factors such as low frequency, as in the case of English (see Pinker, Lebeaux, and Frost 1987). Ultimately, this type of information is critical for developing a comprehensive theory of acquisition.

Spontaneous production studies can also provide information about individual variation in the course of language development. For example, Brown (1973) found that Eve was very precocious in learning the grammatical morphology of English, whereas Adam and Sarah were much slower. This provides researchers with an idea of the range of what can be considered "normal" in language development, and the time course over which it occurs. Although there is thought to be no direct implicational relationship between input and the course of individual children's linguistic development (e.g., Brown 1973), other studies indicate that certain connections may exist. For example, Peters and Menn (1993) argue that the emergence of certain English prepositions in the early speech of two children is closely related to the different input they receive from their respective parents.

Thus, spontaneous production data can provide information regarding the overall course of language development, language-specific and family-specific aspects of the input, individual variation in the developmental

path, and the discourse situations in which language learning takes place. One of the great advantages of collecting and using such data is that they can continue to provide an invaluable source of information regarding various morphological, syntactic, and semantic phenomena as new theoretical questions arise. This is readily attested by the frequent use of Brown's (1973) corpus and others in the CHILDES data bank. In addition, these corpora can provide much-needed information for designing experimental tasks to further tap aspects of linguistic competence. For example, children tend to use restrictive relative clauses in spontaneous speech, yet until Hamburger and Crain 1982, relative clause studies rarely used this type of context in testing children's ability to comprehend and produce relative clauses. As both statistical methods for examining linguistic corpora and connectionist models of learning become more sophisticated, the use of spontaneous production corpora will assume an even greater importance in addressing issues of how syntactic structures are acquired.

1.7 Conclusion

In conclusion, the use of spontaneous production data has had an enormous impact on the field of language acquisition. When carefully collected and coded, these data provide an extremely rich resource for investigating the nature of children's grammatical competence and are invaluable for evaluating hypotheses regarding the acquisition of syntax. The collection of new corpora continues today as emerging theoretical issues call for more data from a larger number of children and a wider range of languages. Recent advances in computer technology, plus the organizational efforts of researchers involved with the CHILDES data archive, provide affordable and widespread access for the use of existing corpora, as well as support for collecting, transcribing, and coding new data sets. These developments lay the groundwork for the continuing importance of spontaneous production corpora for the field of language acquisition.

Note

I thank Shanley Allen, Cecile McKee, and Clifton Pye for comments and discussion.