

Demuth, K. 1996. Collecting spontaneous production data. In D. McDaniel, C. McKee, and H. S. Cairns (eds.), *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press. pp. 3-22.

Collecting Spontaneous Production Data*

Katherine Demuth

Brown University

1. Introduction

Many of the earliest studies of child language acquisition were collected in the form of longitudinal diary studies, where parents documented new developments in their child's developing grammar and/or lexicon (e.g. Stern & Stern 1907, Grégoire 1937, 1947).

Later, with the development of tape recording technology, both parents and non-parent 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 data that could be collected, as well as the types of research questions that could be addressed. Many of these questions, 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 of language acquisition, but also by learning theorist 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 language is actually acquired. Some of the earliest research on the acquisition of English used spontaneous production data to begin to address these questions (e.g. Braine 1963, Brown & Fraser 1963, Miller & Ervin 1964, Bloom 1970). It was also recognized that crosslinguistic data were essential for understanding the nature of the language acquisition process. This lead Slobin and colleagues to the development of *A Field Manual for Cross-Cultural Study of the*

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

Acquisition of Communicative Competence (Slobin 1967). Several studies of children learning other languages followed (Finnish - Bowerman 1967/1973, Samoan - Kernan 1969, and Japanese - McNeill 1966, McNeill & McNeill 1966). Since that time, the collection of spontaneous production data for addressing acquisition questions has become a frequently used methodological tool, and the number of crosslinguistic studies using this technique has continued to grow (e.g. Slobin 1985, 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-bank archive at Carnegie Mellon University (MacWhinney & Snow 1985, MacWhinney 1991). The collection of spontaneous data sets has the potential to make a lasting contribution to the field of language acquisition research. It is not, however, an exercise to be undertaken lightly: Spontaneous production data are only useful when collected systematically and with careful attention to details which affect the quality of the resulting corpus.

One of the sets 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 morpho-syntactically interesting period when the mean length of utterance (MLU) was between 1.75 and 4 morphemes. Although the specific goal of the Brown 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 which can be used to investigate a large number of syntactic issues. For example, the Brown data have been used by Stromswold (1990) in investigating children's acquisition of auxiliary verbs, by Marcus, Pinker, Ullman, Hollander, Rosen, & Xu (1992) in examining

morphological overgeneralization, and by Bloom (1990) in a treatment of children's subjectless sentences. When collected in the appropriate way, spontaneous production data can provide a wealth of information to be taped repeatedly over the years. In the following section I discuss the types of syntactic phenomena that can be most profitably examined using spontaneous production data.

2. Syntactic Phenomena Investigated

One of the primary goals of language acquisition research has been to assess the Chomskian notion of grammatical competence. It is often more difficult to assess young children's knowledge of language than that of adults. Researchers have therefore devised various methodologies appropriate for assessing young children's early grammatical abilities, and many of these are discussed in the later chapters of this book (see McDaniel and Cairns, this volume). Spontaneous production data can also be used to determine certain types of grammatical competence, especially in the area of morpho-syntactic development.

2.1 Pro-drop and Parameter Setting

Since the early and mid 1980's grammatical morphology has played an increasingly important role in the construction of syntactic theory. This has subsequently been reflected in the questions researchers have asked about the course of language acquisition. For example, in the development of Principles and Parameters approach to linguistic structure (Chomsky 1981), it was noted that some languages (e.g. English) have obligatory subjects, whereas others (e.g. Italian) do 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 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), providing statistics on how frequently young English speakers use lexical and pronominal subjects.

2.2 Functional Categories and Syntactic Structure

The relevance of grammatical morphology and its role in children's developing grammars has taken on renewed importance with the incorporation of the distinction between functional vs. lexical categories (closed vs. open class items) into the mainstream of syntactic theory (Abney 1987, Chomsky 1989). A flurry of research activity has ensued on languages as diverse as Italian, English, Swedish, German, Swiss-German, French, Korean, and Sesotho (cf. Meisel 1992, Lust, Suñer, and Whitman 1994, Hoekstra & Schwartz 1994, and references therein), with examination of previously collected spontaneous speech data, as well as the collection of new spontaneous production corpora. These studies have examined how and when various aspects of grammatical morphology are acquired, 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 the Brown 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 researcher such as Pierce (1992) and Déprez & Pierce (1993) to investigate negation in French and the early position of subjects, and by Radford (1994) to examine issues relating to 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).

2.3 Passives, Causatives, and Grammatical Relations

Spontaneous production data have also been used to explore issues of how and when passives, causatives, and other grammatical function changing operations are acquired. Although very few passives are found in spontaneous English, passives are much more commonly found in spontaneous data from children learning Bantu languages (Sesotho - Demuth 1989, 1990, Zulu - Suzman 1985). Children also use ergative marking and anti-passive constructions quite early in languages such as K'iche' (Pye 1992) and Inuktitut (Allen & Crago 1993). Such studies 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), sheds light on children's developing lexicon and lexical interactions with syntactic development. Much of this research uses longitudinal diary studies of children's spontaneous productions, including overgeneralization errors.

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 & McClelland 1986, Pinker & 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, Rothweiler, Woest, & Marcus 1992, Ziesler & Demuth 1994).

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 developments in the area of 'corpus based' linguistics grows, so will the advantages in using spontaneous production data to address acquisition and learnability issues.

3. Spontaneous Production Data Collection Procedures

Spontaneous production data collection, like any other type of data collection, is only useful if collection methods are carefully planned. Planning must include consideration of both the research questions to be asked, as well as 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 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.

3.1 Number of Children to Include in a Study

Acquisition studies have shown that there is a certain amount of individual variation amongst children in the course of language development. While much of this variation is found in the timing of when certain constructions are acquired rather than the course taken to get there, it is nonetheless generally accepted that a study of several children can tell us more than a case 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 choose to discontinue involvement in 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 our construction of a coherent theory of acquisition.

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, there should also be allowance for a certain amount of variation in the age of the children studied: The assessment of individual children's mean length of utterance (MLU) may be a more accurate measure of linguistic ability than age (Brown 1973). This is despite the fact that there may be difficulty calculating MLU crosslinguistically, especially for highly inflected languages such as Hebrew (Dromi & Berman 1982) or West Greenlandic (Fortescue 1985, Fortescue & 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 one is interested in the development of grammatical morphology, it is advisable to begin recordings with children younger than two years in order to catch the transition stage. On the other hand, if grammatical constructions such as passives, relative clauses, or complementation are to be examined, a study that includes older children, perhaps between 2;6 and 4 years, is probably advisable. If a study is designed to look 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 the study.

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.

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 a bilingual or multilingual one, the language situation in the home and/or day-care center should be carefully assessed to insure 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 amongst the children included in the study: This allows for sex-based rates of maturation and gender-based use of language (in some cultures) to be represented. Third, children with a history of health and/or ear infection 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 effect 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 visitations are useful in two respects: First, they offer the researcher an opportunity to become familiar with some

families and their children. Second, they provide a basis for deciding which children will eventually become part of the study. If MLU is a factor in selecting children for the study, it is this period of familiarization which can facilitate assessment of children's stage of linguistic development. Finally, it is not simply the children, but also their families who 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 playing a participatory role in the research. 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. Pre-recording visits to families therefore provide the researcher with critically needed 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 effected if recording sessions are stilted or artificially constructed in any way (cf. Clark 1982).

3.4 Frequency and Duration of Recording Sessions

A decision must be made about the frequency and duration of recording sessions. In studies such as Brown (1973), Eve was recorded for at least half an hour every week, while Adam and Sarah were recorded for around an hour every two weeks. In cases where rapid morpho-syntactic changes were taking place in the child's grammar, more data was collected more frequently. On the other hand, 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 therefore useful to have a plan for how often and how long to record, but also 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 utterances 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 data to collect is to ensure that it constitute a ‘representative sample’ of the child’s productive language capabilities at the time. What counts as ‘representative’ will be highly dependent on the grammatical phenomena being studied and how frequently these constructions occur in everyday discourse. For example, more data is needed for the examination of complex grammatical constructions such as passives, relative clauses, and complementation, whereas less data is needed to examine the use of subject agreement or other frequently-occurring morpho-syntactic 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.

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 to mother-child interactions, nor 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 the child had to be extremely linguistically sophisticated to get what he or she wanted (Demuth 1984). Thus, while the home environment may be the site in which children feel most comfortable, the make up of that home environment may include many more discourse participants than simply the mother. This may be especially true when extended families or peers are living in near proximity, or where the child has older siblings. Such interactions can provide for an extremely rich set of production data, from both the child, as well as other caregivers, including fathers, aunts and uncles, grandmothers, 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 come from a range of daily activities including bathing, cooking, eating, and playing outdoors. Noise factors, such as water running into a bath tub, the TV, washing machine, or

dishwasher in the background, rain pelting on a tin roof, cooking noises, loud music from next door, or ten pre-schoolers 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 therefore 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 at the family's summer camp.

The picture that begins to emerge here is one where the researcher gradually becomes 'part of the extended family'. Both researcher and family will have to make decisions about how this relationship will be negotiated, but 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 daily interactive situation with three families, becoming a member of the extended community and someone the children frequently saw and interacted with. My transitions into and out of families' homes, either with or without the tape recorder, became normal events in the life of each child, allowing me to record whenever and where ever the collection of spontaneous linguistic productions looked promising. Sometimes this meant joining families for a meal, at 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.

One decision the researcher will have to make is 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 limit his or her 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 the interpretation of the child 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 video tape, that may provide the information needed to evaluate children's use of such constructions. Written or video-taped contextual notes should be keyed to the counter on the tape recorder for easy and accurate retrieval.

3.6 Recording Equipment

The collection of spontaneous production data may take place 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 you will be recording, is highly recommended. This is especially relevant as current technologies and the recording products available continue to change. 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 going to be 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 a large one with the potential for being of significant long-term interest to the field, as in the case of the Brown (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 disappearing to the field. This is especially true with the development of new technologies, where the researcher may need to become familiar with the equipment, or where 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 selection of the tape recorder is the choice of the microphone. Again, aspects of the recording site will factor largely 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 voices from all speakers can be heard. However, if recordings will take place moving from one site to another, it is advisable to use either a hand held directional microphone, or a wireless broadcast microphone built into a vest that the child wears, or both if discourse interactions with other speakers is also desired. The quality of the microphones should be good. If recording is done outdoors, microphones should be used with a wind screen to reduce noise from wind.

Given the nature of collecting spontaneous production data from young children, it may be best to run the recorder off of 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, it is advisable at the end of each recording session to verify that recording actually took place, and that it is intelligible. All tapes should be marked with the recording date and age and name of the child.

4. Transcribing and Tagging Spontaneous Production Data

Several recent publications have dealt the transcription and tagging procedures most appropriate for different types of production data (cf. *The CHILDES Project: Tools for Analyzing Talk* - MacWhinney 1991, and *Talking Data: Transcription and Coding in Discourse Research* - Edwards & Lampert 1993). The reader is referred to those sources for information on widely used conventions for transcribing and tagging spontaneous speech data. The purpose of this section is not to reiterate the material found there, but rather to provide a procedural perspective on these issues, with specific reference to the types of decisions that will need to be made.

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 carry out transcription on the copy: 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, as most transcribing machines still use analog tape. Transcription

can then be carried out on the copy tape, with the original kept for archival use, or as a backup if needed.

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 Stromswold, 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 examples of databases which 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, including the child's name (using a pre-selected pseudonym), the child's age, the date and site of the recording, and participants in the recording session (e.g. mother, other relatives or siblings, friends). If the researcher is not a native speaker of the language being investigated, it may be advisable to conduct the transcription process 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 session.

Finally, the researcher will have to decide whether to transcribe only the child's utterances, or whether to include both 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.

4.2 The Transcription Process

As Ochs (1979) so aptly noted, the very process of transcription has theoretical consequences. At every stage of the data collection and transcription process, certain details are lost, resulting in an end product that preserves only certain types of information. Given that researchers generally use only the final transcript when conducting syntactic analysis (and when using transcripts in the CHILDES data bank this 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 non-laboratory 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 as to 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 which use lexical and/or grammatical tone, such as many south-east Asian and African languages, tone may need to be marked to ensure lexical, morphological, or syntactic information. The presence or absence of grammatical function items, including vowel or consonantal 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, to the extent possible, even a broad phonemic transcription may need to be carried out with attention to some phonetic detail. In some cases, especially with languages other than English, this may involve the use of diacritics not readily available on an English keyboard. The CHILDES manual (MacWhinney 1991) has a series of conventions generally used for 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 concerns 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. Once in database format, it is also possible to recover the complete utterances so that relative clauses and embedded complements can also be examined. On the other hand, Allen (1994) transcribed one utterance per 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 have generally coded identical consecutive repetitions as one utterance, with a note in the ‘comments’ column (or ‘field’ in the data base) as to how many times it occurred. If ‘repetitions’ are segmentally or prosodically different, or if they occur with intervening speech from other interlocutors, I count these 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, cleft constructions, relativization, the use of stressed pronouns, word order, and the like.

Contextual information that has been captured in notes or on video tape should therefore be entered into a 'context' field in the transcript.

Even with the aid of contextual information, a video tape, and the assistance of a speaker of the language, it may occasionally be difficult to determine what the child actually said. Sometimes nothing or very little 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 make notes in a 'comments' field of the transcript that there is some uncertainty about what has been transcribed, and include alternatives if there are any. Some of these utterances may become disambiguated once more transcription has been completed. 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 transcription has been completed, it should be checked/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 time consuming process. I found that, even working with the mother or grandmother of the child, broad phonemic transcription of the Sesotho corpus generally took seven hours for one hour of audio tape. Allen (1994) and Crago (1988) found that Inuktitut speakers transcribed video tape at the rate of about one hour for 2-5 minutes of video tape. In other words, transcribing either audio or video tape requires a large investment of time. The researcher should plan accordingly.

4.3 Tagging (Coding) the Corpus

Spontaneous production corpora are most useful if some type of tagging (grammatical coding) is included in the transcript (data base). 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 essential when working with a lesser known language - especially if the eventual goal is to make the corpus available to a larger audience, such as 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 gloss) 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	x2

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

(2)	1sSM	1st person singular subject agreement
	2sOM	2nd person singular object agreement
	PRES	present tense
	IN	indicative mood
	x2	identical consecutive repetition of an utterance

The specific tags used will be partially dependent on the language under investigation and partly dependent on the research questions being asked. This type of detailed tagging system is extremely useful for conducting automatic searches of certain grammatical phenomena, especially in cases where children's pronunciation 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. This information is often lacking in both transcripts and in publications, making it difficult for both researchers and readers who do not have full command of particular language to understand what the child has omitted or changed.

5. Disadvantages of Collecting/Using Spontaneous Production Data

Some of the potential problems involved with collecting spontaneous production data have already been discussed in section 4. Once collected, however, there are also certain limitations on what the data 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 therefore introduces certain problems of interpretation: How and when does the researcher know that the child has *productive use* of certain grammatical forms? These issues are discussed below.

One of the limitations of using spontaneous production data is the nature of the sampling technique itself: If a particular grammatical construction does not arise 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 this lack of passives in spontaneous production data was due to grammatical complexity or lack of linguistic ‘maturation’ (e.g. Brown & Hanlon 1970, Borer & 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 is of limited use (without crosslinguistic evidence) in determining if 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 a some grammatical constructions, especially those pertaining to grammatical morphology, may precede children’s actual production of these forms (e.g. Shipely, Smith, & Gleitman 1969). This is especially relevant for the current debate concerning the presence or absence of functional categories and the projection of syntactic structure (cf. Meisel 1992, Lust, Suñer, & Whitman 1994, Hoekstra & 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 only limited control of relative clauses (e.g. de Villiers, Tager-Flusberg, Hakuta & Cohen 1979, Tavakolian 1981) and long distance wh-movement (de Villiers, Roeper & 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 the use of morphological ‘errors’ such as the overgeneralization of past tense *-ed*. (e.g. *goed, catched*). 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.

6. Advantages of Collecting/Using Spontaneous Production Data

The greatest advantage of using spontaneous production data is that it can supply a wealth of information about many aspects of children’s grammatical development. When spontaneous production data is carefully collected, transcribed and tagged for several children over a long period of time, the resulting corpus provides an invaluable record of children’s developing linguistic competence that can be used to address any number of theoretical issues.

Longitudinal spontaneous production studies are particularly useful in identifying general developmental trends, providing an excellent picture of the overall course of development for a particular language. This is especially useful when initiating the study of a language for which there has been little or no previous acquisition research. 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 & Quixtan Poz 1988) show early acquisition of anti-passive 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 which 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 are especially useful in providing information about the frequency with which specific grammatical constructions typically occur in a language. They can therefore provide important evidence for determining if the acquisition of a particular grammatical phenomenon, such as the passive, is linguistically difficult for young children, or simply fails to appear due to language particular discourse factors such as low frequency, as in the case of English (cf. Pinker, Lebeaux, & 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 children like Eve are very precocious in learning the grammatical morphology of English, while children like Adam and Sarah are 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 a direct implicational relationship between input and the course of individual children's linguistic development is thought to be missing (e.g. Brown 1973), other studies indicate that there may be certain connections. For example, Peters & 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, individual variation in that developmental path, and the discourse situations in which language learning takes place. One of the great advantages of collecting and using this type of data is that it 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 the Brown (1973) corpus and others in the CHILDES data bank. In addition, these corpora can also provide much needed background for refining further research questions, including the construction of 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 the use of 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 syntax is acquired.

7. Conclusion

In conclusion, the collection and use of spontaneous production data has had an enormous impact on the field of language acquisition. These data provide an extremely rich resource for investigating the nature of children's grammatical competence, and have become invaluable for evaluating hypotheses regarding the acquisition of syntax. The collection of new corpora continues today as new theoretical issues call for more data from a larger range of children and languages. Recent advances in computer technology,

plus the organizational efforts of researchers involved with the CHILDES data bank, 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.

References

- Abney, S. (1987). The English noun phrase in its sentential aspect. Doctoral dissertation, MIT, Cambridge.
- Allen, S. (1994). Acquisition of some mechanisms of transitivity alternation in arctic Quebec Inuktitut. Doctoral dissertation, McGill University, Montreal, QC.
- Allen, S., & Crago, M. (1993). Early acquisition of passive morphology in Inuktitut. Proceedings of the 24th Annual Child Language Research Forum, (pp. 112-123). Stanford: CSLI.
- Bellugi, U. (1967). The acquisition of negation. Doctoral dissertation, Harvard University, Cambridge, MA.
- Bloom, L. (1970). Language development: Form and function in emerging grammars. Cambridge: MIT Press.
- Bloom, P. (1990). Subjectless sentences in child language. Linguistic Inquiry, 21, 491-504.
- Borer, H., & Wexler, K. (1987). The maturation of syntax. In T. Roeper & E. Williams (Eds.), Parameter-setting and language acquisition (pp. 123-172). Dordrecht: D. Reidel Publishing Company.
- Bowerman, M. (1973). Early syntactic development: A cross-linguistic study with special reference to Finnish. Cambridge : Cambridge University Press.

- Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner & L. Gleitman (Eds.), Language acquisition: The state of the art. Cambridge, England, Cambridge University Press.
- Braine, M. (1963). The ontogeny of English phrase structure: The first phase. Language, 39, 1-13.
- Brown, R. (1968). The development of wh questions in child speech. Journal of Verbal Learning and Verbal Behavior, 7, 279-290.
- Brown, R. (1973). A first language. Cambridge, MA: Harvard University Press.
- Brown, R., & Fraser, C. (1963). The acquisition of syntax. In C. N. Cofer & B. Musgrave (Eds.), Verbal behavior and learning: Problems and processes. New York: McGraw-Hill.
- Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), Cognition and the development of language. New York: Wiley.
- Chomsky, N. (1957). Syntactic structures. The Hague: Mouton.
- Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.
- Chomsky, N. (1981). Lectures on government and binding. Dordrecht, Foris Publications.
- Chomsky, N. (1989). Some notes on the economy of derivation and representation. MIT Working Papers in Linguistics 10, 43-74.
- Clark, R. (1982). Theory and method in child-language research: Are we assuming too much? In S. Kuczaj, II (Ed.), Language development volume 1: Syntax and semantics (pp. 1-36). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Clahsen, H., Rothweiler, M., Woest, A. & Marcus, G. F. (1992). Regular and irregular inflection in the acquisition of German noun plurals. Cognition, 45, 225-255.
- Crago, M. B. (1988). Cultural context in communicative interaction of Inuit children. Doctoral dissertation, McGill University, Montreal.

- Demuth, K. (1984). Aspects of Sesotho language acquisition. Indiana University Linguistics Club, Bloomington.
- Demuth, K. (1989). Maturation and the acquisition of Sesotho passive. Language, *65*, 56-80.
- Demuth, K. (1990). Subject, topic and the Sesotho passive. Journal of Child Language, *17*, 67-84.
- Demuth, K. (1992). The acquisition of Sesotho. In D. I. Slobin (Ed.), The Cross-Linguistic Study of Language Acquisition, Vol 3 (pp. 557-638). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Demuth, K. (1993). Issues in the Acquisition of the Sesotho Tonal System. Journal of Child Language, *20*, 275-301.
- Demuth, K. (1995). Questions, Relatives, and Minimal Projection. Language Acquisition, *4*, 49-71.
- Déprez, V., & Pierce, A. (1993). A crosslinguistic study of negation and functional projections in early grammar. Linguistic Inquiry, *24*, 25-67.
- de Villiers, J., Roeper, T., & Vainikka, A. (1990). The acquisition of long-distance rules. In L. Frazier & J. de Villiers (Eds.), Language processing and language acquisition. Dordrecht: Kluwer Academic Publishers.
- de Villiers, J., Tager-Flusberg, H., K. Hakuta, K., & Cohen, M. (1979) Children's comprehension of the relative clause. Journal of Psycholinguistic Research, *8*, 499-518.
- Dromi, E., & Berman, R. (1982). A morphemic measure of early language development: data from modern Hebrew. Journal of Child Language, *9*, 403-424.
- Edwards, J., & Lampert, M. (1993). Talking data: Transcription and coding in discourse research. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Fortescue, M. D. (1985). Learning to speak Greenlandic: A case study of a two-year-olds' morphology in a polysynthetic language. First Language, *5*, 101-114.

- Fortescue, M. D., & Lennert Olsen, L. (1992). The acquisition of West Greenlandic. In D. I. Slobin (Ed.), The cross-linguistic study of language acquisition, Vol 3 (pp. 111-220). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Grégoire, A. (1937). L'apprentissage du langage. Les deux premières années. Liège/Paris. (no place of publication).
- Grégoire, A. (1947). L'apprentissage du langage. Les troisième année et les années suivantes. (no place of publication).
- Hamburger, H. & Crain, S. (1992). Relative acquisition. In S. Kuczaj (Ed.), Language Development: Syntax and Semantics. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hoekstra, T. & Schwartz, B. (1994). Language acquisition studies in generative grammar. Amsterdam: John Benjamins.
- Hyams, N. (1986). Language acquisition and the theory of parameters. Dordrecht: Reidel.
- Kernan, K. T. (1969). The acquisition of language by Samoan children. Doctoral dissertation, University of California at Berkeley, Berkeley.
- Lust, B., Suñer, M., and Whitman, J., (Eds.). (1994). *Syntactic theory and first language acquisition: Crosslinguistic perspectives*, 1, Hillsdale, N.J., Lawrence Erlbaum Associates.
- MacWhinney, B. (1991). The CHILDES Project: Tools for analyzing talk. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- MacWhinney, B., & Snow, C. (1985). The Child Language Data Exchange System. Journal of Child Language, 12, 271-296.
- Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). overregularization in language acquisition. Monographs of the Society for Research in Child Development, 57. (No. 4, Serial No. 228).

- McNeill, D. (1966). The creation of language by children. In J. Lyons & R. J. Wales (Eds.), Psycholinguistic papers (pp. 99-115). Edinburgh: Edinburgh University Press.
- McNeill, D., & McNeill, N. (1966). What does a child mean when he says “no”? In E. Sale (Ed.), Proceedings of the conference on language and language behavior (pp. 51-61). New York: Appleton-Century-Crofts.
- Meisel, J. (Ed.). (1992). The acquisition of verb placement: Functional categories and V2 phenomena in language development. Dordrecht: Kluwer Academic Publishers.
- Miller, W., & Ervin, S. (1964). The development of grammar in child language. in U. Bellugi & R. Brown (Eds.), The acquisition of language. Monographs of the Society of Research in Child Development, 29, 9-34.
- Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), Developmental Pragmatics (pp. 43-72). New York: Academic Press.
- Pierce, A. (1989). Language acquisition and syntactic theory. Dordrecht: Kluwer Academic Publishers.
- Pinker, S., Lebeaux, D., & Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. Cognition, 26, 195-267.
- Peters, A., & Menn, L. (1993). False starts and filler syllables: Ways to learn grammatical morphemes. Language, 69, 742-778.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition, 28, 73-193.
- Pye, C. (1992). The acquisition of K'iche' Maya. In D. I. Slobin (Ed.), The cross-linguistic study of language acquisition, Vol 3 (pp. 221-308). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Pye, C., & Quixtan Poz, P. (1988). Precocious passives (and antipassives) in Quiche Mayan. Papers and Reports on Child Language Development, 27, 71-80.

- Radford, A. (1994). The syntax of questions in child English. Journal of Child Language 21, 211-236.
- Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. Implicit rules or parallel distributed processing? In J. McClelland, D. Rumelhart, and the PDP Research Group, Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press.
- Shipley, E., Smith C., & Glietman, L. (1969). A study in the acquisition of language: Free responses to commands. Language, 45, 322-342.
- Slobin, D. I. (Ed.). (1967). A field manual for cross-cultural study of the acquisition of communicative competence. Berkeley: ASUC Bookstore.
- Slobin, D. I. (Ed.). (1985). The cross-linguistic study of language acquisition, Vol 1 & 2. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Slobin, D. I. (Ed.). (1992). The cross-linguistic study of language acquisition, Vol 3. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Stern, C., & Stern, W. (1907). Die Kindersprache: Eine psychologische und sprachtheoretische Untersuchung. Leipzig: Barth.
- Stromswold, K. J. (1990). Learnability and the acquisition of auxiliaries. Doctoral dissertation, MIT, Cambridge, MA.
- Suzman, S. (1985). Learning the passive in Zulu. Papers and Reports on Child Language Development, 24, 131-37. Stanford: Stanford University.
- Tavakolian, S. (1981) The conjoined-clause analysis of relative clauses. In S. Tavakolian, ed., Language Acquisition and Linguistic Theory, Cambridge, MA, MIT Press.
- Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. Cognition, 40, 21-81.

Ziesler, Y., & Demuth, K. (1994). Noun class prefixes in Sesotho child-directed speech.

In E. Clark (Ed.), Proceedings of the 26th Annual Child Language Research

Forum (pp. 137-146). Stanford: CSLI.