Corpus linguistics and language acquisition

Holger Diessel


Address:
Holger Diessel
Institut für Anglistik/Amerikanistik
Friedrich-Schiller-Universität Jena
Ernst-Abbe-Platz 8
07743 Jena
Phone: 03641/944530
Email: holger.diessel@uni-jena.de
Outline

1. Introduction

2. Linguistic corpora in child language research
   2.1. A brief history of the study of spontaneous child speech
   2.2. The CHILDES database
   2.3. The CHILDES computer programs

3. Sample studies
   2.1. The acquisition of grammatical categories
   2.2. The acquisition of complex sentences

4. Conclusion

References
1. Introduction

Child language researchers use two basic methodological approaches to the study of language acquisition: They use experiments to test children’s linguistic knowledge in controlled situations, and they collect spontaneous child language data to analyze their linguistic behaviour in natural settings (cf. McDaniel et al. 1996). The two approaches complement each other: Experimental methods are primarily used to test particular hypotheses, and observational methods are commonly used to explore the development of new aspect or phenomenon. When observational data are transcribed and compiled to a large database they constitute a linguistic corpus. This chapter is concerned with linguistic corpora in the field of language acquisition. It is divided into two parts. The first part describes the design and history of child language corpora, notably the CHILDES database, and considers some of the research tools that child language researchers use to analyze spontaneous child language data; and the second part discusses two recent sample studies that have drawn on data from the CHILDES database.
2. Corpus-linguistic approaches to the study of language acquisition

2.1. A brief history of the study of spontaneous child speech

Today child language corpora are computerized and pre-processed by automatic taggers, but the study of spontaneous child language started long before the advent of computers and modern corpus linguistics. The first investigations of natural child language appeared more than a hundred years ago; they were parental diary studies tracing the linguistic development of a particular child based on the parents’ observations. One of the earliest and most frequently cited diary studies was a study by Clara and William Stern (1907). The Sterns carefully documented the speech of their two German-speaking children, Hilde and Gunter, and proposed the first widely accepted stages of child language development. Another early diary study that had an important impact on the field was an investigation by Leopold (1939-49), who described the linguistic development of his daughter Hildegard from birth to age two.

In the 1940\textsuperscript{th} and 1950\textsuperscript{th}, behaviourist psychologists began to collect child language data more systematically. In contrast to the early diary studies, the behaviourists worked primarily with cross-sectional data; that is, rather than studying the development of one child in detail, they collected data from a large number of children using small language samples of about an hour of speech. Some of these cross-sectional investigations included data from more than 100
hundred children carefully selected based on their ages, sexes, and socio-economic backgrounds.

Starting in the 1960\textsuperscript{th}, the first systematic longitudinal studies appeared. Like the early diary studies, these studies traced the development of a particular child (or a few children) over an extended period of time. However, in contrast to the early diary studies, these studies did not use parental observations as their primary source; rather, they used a new technology to collect naturalistic data. For the first time in history, child language was first audio-taped and then carefully transcribed and annotated by the researcher. While parental diary data is fairly reliable to investigate children’s early one- and two-word utterances, it is impossible to keep a systematic diary of more advanced children. The use of audio tapes allowed the researcher to systematically document and analyze more complex linguistic phenomena that emerge only after the two-word stage. The studies that appeared at this time include the pioneering works by Martin Braine (1963, 1976), Lois Bloom (1970), and Roger Brown (1973), which have become classics of child language research.

During the 1970\textsuperscript{th} and 1980\textsuperscript{th}, the number of longitudinal studies increased steadily; however, the data collected were only available to a small number of researchers (usually the researchers who had collected the data and their colleagues). Advances in computer technology made it possible to share child language data more easily. Starting in the early 1980\textsuperscript{th}, Brian MacWhinney and Catherine Snow established a large computerized database, which they called the
Child Language Exchange System, or short CHILDES (cf. MacWhinney and Snow 1985). The CHILDES database includes data from more than a hundred research projects from a variety of languages. MacWhinney and Snow compiled these data and developed a number of research tools to facilitate their analysis. Among other things, they developed a coding system to standardize the transcriptions and a set of computer programs for searching and manipulating the transcripts. Both the database and the computer programs are freely available on the internet. Over the years, CHILDES has grown to a whole research system, providing not only a comprehensive database and computer programs, but also access to an electronic bibliography, a mailing list, and various other research tools (http://childes.psy.cmu.edu/). The two following sections describe the database and some of the research tools.

**The CHILDES database**

The CHILDES database is very different from corpora of adult corpus linguistics. In contrast, to adult corpora such as the British National Corpus (BNC) or the International Corpus of English (ICE), the CHILDES database does not have a pre-determined structure or a specific design. It includes data from many different research projects that researchers have made available to the scientific community. By contrast, adult corpora are designed based on a pre-determined set of criteria. For instance, the British National Corpus was designed
such that it includes data from many different varieties of English. It contains both written texts and transcripts of spoken discourse that represent a wide variety of different genres and speech registers. In contrast, the CHILDES database contains only spoken data. Most of these data come from naturally occurring conversations between the child and an adult speaker, notably one of the child’s parents. Written language and different speech registers have been neglected because this type of variation does not exist in early child language. That does not mean that child language is entirely homogenous. There are two dimensions of variation that play an important role in the analysis of language acquisition: differences between individual children and differences between different ages of an individual child. The two dimensions of variation are reflected in the overall structure of the CHILDES database: The data are divided into files of individual children containing transcripts of speech from different ages. Table 1 shows the subdirectory of one child in the CHILDES database. In this case, the subdirectory contains 20 files including transcripts that were prepared at different times between the ages of 1;9 and 3;1.

<table>
<thead>
<tr>
<th>File</th>
<th>Age</th>
<th>File</th>
<th>Age</th>
<th>File</th>
<th>Age</th>
<th>File</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>1:9.7</td>
<td>06</td>
<td>2:0.7</td>
<td>11</td>
<td>2:3.21</td>
<td>16</td>
<td>2:7.14</td>
</tr>
<tr>
<td>02</td>
<td>1:9.21</td>
<td>07</td>
<td>2:0.7</td>
<td>12</td>
<td>2:4.14</td>
<td>17</td>
<td>2:8.14</td>
</tr>
<tr>
<td>03</td>
<td>1:10.15</td>
<td>08</td>
<td>2:1.21</td>
<td>13</td>
<td>2:5.0</td>
<td>18</td>
<td>2:9.14</td>
</tr>
<tr>
<td>04</td>
<td>1:11.7</td>
<td>09</td>
<td>2:2.14</td>
<td>14</td>
<td>2:5.21</td>
<td>19</td>
<td>2:10.21</td>
</tr>
<tr>
<td>05</td>
<td>1:11.21</td>
<td>10</td>
<td>2:3.0</td>
<td>15</td>
<td>2:6.14</td>
<td>20</td>
<td>3:1.21</td>
</tr>
</tbody>
</table>
The subdirectories of other children are similar, but they may cover a different age span and the recordings may have occurred at different intervals. Most of the data consist of one hour recordings that were regularly prepared once or twice a month. The age range differs between several years and a few months. Most of the data come from longitudinal studies of monolingual children, but there is also data from a few cross-sectional studies and data from bilingual children. Moreover, CHILDES includes some data from second language learners, bilingual children, and children with language disorders (e.g. Down syndrome, autism, SLI = Specific Language Impairment) (cf. MacWhinney 2000). Finally, there is data from different languages. The bulk of the data come from children learning English, but apart from English there is data on more than 20 other languages, including data from non-Indo-European languages such Hebrew, Turkish, Hungarian, Japanese, and Mandarin Chinese.

All of the data appear in the CHAT format, which has been designed to accommodate a large variety of levels of analysis. Every CHAT file includes a header providing information about the participants, the situational context, and general information about the file such as the data of the recording and the length of the transcript.

@Begin
@Languages: en
@Participants: CHI Adam Target_Child, MOT Mother, URS Ursula_Bellugi Investigator, FAT Father, PAU Paul Brother
@ID: en|brown|CHI|4;0.14|male|normal|middle_class|Target_Child||
@ID: en|brown|MOT|||||Mother||
@ID: en|brown|URS|||||Investigator||
@ID: en|brown|FAT|||||Father||
@ID: en|brown|PAU|||||Brother||
@Date: 23-JUL-1964
@Time Duration: 10:00-11:00
@Situation: Adam has a new puppy and is very wary of him. When the puppy is out of his enclosure, Adam stays up on his bed.

The actual transcript consists of the main tier exhibiting the speech of the participants and one or more dependent tiers.

*CHI:  Mommy # we have to do something # lock him up.
%mor:  n:prop|Mommy pro|we v:aux|have inf|to v|do pro:indef|something
   v|lock pro|him prep|up.
*MOT:  he won't bother you.
%mor:  pro|he v:aux|will~neg|not v|bother pro|you.
*MOT:  he's all tired out.
%mor:  pro|he~v|be&3S qn|all adj|tired prep|out.
*CHI:  Mommy +...
%mor:  n:prop|Mommy +...
*MOT:  his name is Rinny from Rin_tin_tin.
%mor:  pro:poss:det|his n|name v|be&3S n:prop|Rinny prep|from
    n:prop|Rin_tin_tin
Every utterance is represented by a separate line on the main tier. So if a speaker produces several utterances in one turn, the turn is represented on several lines. There is a set of diacritics that mark the end of an utterance, providing important information for some of the CLAN programs (see below). Apart from the diacritics for utterance boundaries, the main line may include diacritics for pause (#), for incomplete utterances (+…), for false starts, repetitions, omissions, and various other aspects of the utterance.

The main tier may have several dependent tiers. One of the dependent tiers is obligatory. All of the data included in the CHILDES database is tagged and morphologically analyzed. This information is coded on the %mor line, which occurs right underneath the main tier. In addition to the %mor line, an utterance may have several other dependent tiers providing information about the situation context, the intonation, and certain types of errors.

Most of the data in the CHILDES database consist of one hour recordings that were prepared at regular intervals of one to three weeks. Using this sampling method, the transcripts represent between 1.0% to 1.5% of the language children hear and produce during the time of the study (cf. Tomasello and Stahl 2004). While 1.5% of the data may provide enough information to trace the development of frequently occurring structures and expressions, they are not sufficient to study the development of less frequent phenomena. As Tomasello
and Stahl (2004) have shown, in order to trace the development of infrequent grammatical patterns one needs to examine a much larger proportion of children’s speech. In particular, the age of appearance and the order of acquisition are not reliable if we use the standard language samples of one hour recorded every other week; rather, what we need are much denser data collections, as they are currently prepared at the Max Planck Institute for Evolutionary Anthropology in Leipzig.

The CLAN programs
Another important issue that requires careful planning concerns the quantity of the data. How much data is needed in order to investigate the development of a particular phenomenon? Specifically, how much data is needed to determine the age of appearance, the order of acquisition, and the developmental pathway? Until recently, most researchers collected their data in one hour recordings that were prepared at regular intervals of one to three weeks. Using this sampling method, the transcripts represent between 1.0% to 1.5% of the language children hear and produce during the time of the study (cf. Tomasello and Stahl 2004). While 1.5% of the data may provide enough information to trace the development of frequently occurring structures and expressions, they are not sufficient to study the development of less frequent phenomena. As Tomasello and Stahl (2004) have shown, in order to trace the development of infrequent grammatical patterns one needs to examine a much larger proportion of children’s speech. In particular, the age of appearance and the order of acquisition are not reliable if we use the standard language samples of one hour recorded every other week; rather, what we need are much denser data collections, as they are currently prepared at the Max Planck Institute for Evolutionary Anthropology in Leipzig.
2. Collecting and analyzing child language data.

Ideally, every researcher would establish a particular database to investigate a specific research question. However, since preparing a child language corpus is very time consuming and expensive, researchers usually collect their data in such a way that they can be used to address a variety of research questions. How must such a corpus be designed to be designed?

In adult corpus linguistics, a multi-purpose corpus includes data that are representatives of different genres. For instance, the British National Corpus includes both written texts and transcripts of spoken discourse, which have been selected from a variety of different genres and speech registers. A child language corpus consists exclusively of spoken language, which is usually taken from spontaneous conversations between the child and an adult speaker. In other words, a child language corpus does not include data from different genres and speech registers, because this type of variation does not exist in early child speech.

That does not mean that child language is entirely homogenous. There are two important dimensions of variation that play an important role in the study of language acquisition: First there are significant differences in the development of individual children, and second an individual child uses language differently at different ages. These two dimensions determine the overall design of a child language corpus. In contrast to an adult corpus, a
child language corpus is commonly divided into separate files for individual children containing data from different ages.

However, the number and age of children are only two of the factors that are important to the design of a child language corpus. In addition, there are various socio-economic factors that researchers have to take into account. For instance, one factor that can have an important effect on the course of language acquisition is the socio-economic background of the child: there can be important differences between children from different economic classes, from different ethnic backgrounds, or different areas. Another important factor is the sex of the child: there is good evidence that boys and girls tend to acquire language in somewhat different ways.

Another important issue that requires careful planning concerns the quantity of the data. How much data is needed in order to investigate the development of a particular phenomenon? Specifically, how much data is needed to determine the age of appearance, the order of acquisition, and the developmental pathway? Until recently, most researchers collected their data in one hour recordings that were prepared at regular intervals of one to three weeks. Using this sampling method, the transcripts represent between 1.0% to 1.5% of the language children hear and produce during the time of the study (cf. Tomasello and Stahl 2004). While 1.5% of the data may provide enough information to trace the development of frequently occurring structures and expressions, they are not sufficient to study the
development of less frequent phenomena. As Tomasello and Stahl (2004) have shown, in order to trace the development of infrequent grammatical patterns one needs to examine a much larger proportion of children’s speech. In particular, the age of appearance and the order of acquisition are not reliable if we use the standard language samples of one hour recorded every other week; rather, what we need are much denser data collections, as they are currently prepared at the Max Planck Institute for Evolutionary Anthropology in Leipzig.

2.2. From diary studies to the study of large computerized corpora

The study of spontaneous child language has a long history. The first investigations appeared more than a hundred years ago; they were parental diary studies tracing the linguistic development of a particular child based on the parents’ observations. One of the earliest and most frequently cited diary studies was a study by Clara and William Stern (1907). The Sterns carefully documented the speech of their two German-speaking children, Hilde and Gunter, and proposed the first widely accepted stages of child language development. Another early diary study that had an important impact on the field was an investigation by Leopold (1939-49), who described the linguistic development of his daughter Hildegard from birth to age two.
In the 1940\textsuperscript{th} and 1950\textsuperscript{th}, behaviourist psychologists began to collect child language data more systematically. In contrast to the early diary studies, the behaviourists worked primarily with cross-sectional data; that is, rather than studying the development of one child in detail, they collected data from a large number of children using small language samples of about an hour of speech. Some of these cross-sectional investigations included data from more than 100 hundred children carefully selected based on their ages, sexes, and socio-economic backgrounds.

Starting in the 1960\textsuperscript{th}, the first systematic longitudinal studies appeared. Like the early diary studies, these studies traced the development of a particular child (or a few children) over an extended period of time. However, in contrast to the early diary studies, these studies did not use parental observations as their primary data; rather, they used a new technology to collect data. For the first time in history, child language was first audio-taped and then carefully transcribed and annotated by the researcher. While parental diary data is fairly reliable for investigating children’s early one- and two-word utterances, it is impossible to keep a systematic diary of more advanced children. The use of audio tapes allowed the researcher to systematically document and analyze more complex linguistic phenomena that emerge after the two-word stage. The studies that appeared at this time include the pioneering works by Martin Braine (1963,
1976), Lois Bloom (1970), and Roger Brown (1973), which have become classics of child language research.

During the 1970th and 1980th, the number of longitudinal studies increased steadily; however, the data collected were only available to a small number of researchers (usually the researchers who had collected the data and their colleagues). Advances in computer technology made it possible to share child language data more easily. Starting in the early 1980th, Brian MacWhinney and Catherine Snow established a large computerized database, which they called the Child Language Exchange System, or short CHILDES (cf. MacWhinney and Snow 1985).

2.3. CHILDES

CHILDES provides a tremendous research tool for the study of child language. It does not only consist of a large database of spontaneous child speech, but also of a set of computer programs, called CLAN, for searching and manipulating the database. Moreover, MacWhinney and Snow developed a coding system, called CHAT, which helped to standardize the transcriptions included in the database. Today, the CHILDES database and the CLAN programs are freely available on the internet. CHILDES has grown to a whole research system, providing not only a comprehensive database and computer programs, but also access to an electronic
bibliography, a mailing list, and various other research tools (http://childes.psy.cmu.edu/).

The database consists of transcripts from over sixty projects on child language development in English and data from more than twenty other projects on the acquisition of other languages, including non-Indoeuropean languages such as Turkish, Mandarin, and Japanese. In addition, the CHILDES database includes data from second language learners, bilingual children, and children with language disorders (e.g. Down syndrome, autism, SLI = Specific Language Impairment) (cf. MacWhinney 2000).

All of the data appear in the CHAT format, in which utterances are accompanied by several dependent tiers, providing information on the discourse context, nonverbal communication, illocutionary force, intonation, and grammar. Most of the English data have been automatically tagged and morphologically analyzed. This information is coded on the %mor line, which also provides information on certain types of errors. For instance, if the child does not produce a grammatical marker that is obligatory in adult language (error of omission) or if he/she uses a grammatical marker that is not permissible in adult language (error of commission), this is indicated on the %mor line (cf. MacWhinney 1995).
The CLAN computer programs have been specifically designed to analyze child language data. For instance, there is a program that allows the researcher to calculate the MLU, i.e. the Mean Length of Utterance. The MLU is a particular measure, suggested by Roger Brown (1973), to determine the child’s linguistic knowledge. Since there are significant differences in the linguistic development of individual children, the MLU is a much more reliable measure than the child’s age to indicate the stage of linguistic development. Apart from the MLU program, CLAN includes programs for searching words and collocations, for determining type and token frequencies, for analyzing phonemes, and several other tasks (cf. MacWhinney 2000).

Some of the most recent contributions to the CHILDES database consist of transcripts that are linked to audio- and video-files. These multi-media files allow the researcher to systematically investigate new aspects of child language. For instance, although some of the standard CHAT transcripts include information on intonation and illocutionary force, it is impossible to systematically investigate these aspects without the linkage of transcripts and audio-recordings. Similarly, the linkage of transcripts and video-files makes it possible to study non-verbal communication and the influence of the pragmatic context on the child’s speech.

1.2. Collecting and analyzing spontaneous child language
Ideally, every researcher would establish a particular database to investigate a specific research question. However, since preparing a child language corpus is very time consuming and expensive, researchers usually collect and transcribe their data in such a way that they can be used to address a variety of research questions. Most of the data included in the CHILDES database document the development of a particular child without focusing on a specific issue. However, the transcripts have been primarily prepared to investigate the development of lexical and morphosyntax phenomena. The recordings are transcribed based on orthographic conventions of English (or another language); phonetically altered forms are only marked if they are particularly striking and/or if the altered forms are well-known from adult language (e.g. ‘wanna’ instead of ‘want to’). Intonational and contextual cues are marked by diacritics if they are relevant to the interpretation of the conversation, but these diacritics are not sufficient to investigate intonational or kinetic aspects of early child speech. In order to study these aspects, one has to use other types of data (notably transcripts linked to audio- and video-files).

With a few exceptions, all of the data included in the CHILDES database are data from longitudinal studies. However, in order to address a particular research question, the researcher may use a subset of the data from individual children to build a cross-sectional corpus. Most recent investigations do not just study one child but draw on data from several
children even if they do not employ an explicit cross-sectional design. In general, the more children are examined, the more representative the investigation.

However, the number of children is only one of the factors that a researcher has to take into account in order to compile a representative language sample. Equally important are the age, sex and the socio-economic background of the children. Most of the children included in the CHILDES database come from highly educated middle-class families. A number of researchers have pointed out that the linguistic development of these children may be different from the development of children of other social groups.

Another important issue that requires careful planning concerns the quantity of the data. How much data is needed in order to investigate the development of a particular phenomenon? Specifically, how much data is needed to determine the age of appearance, the order of acquisition, and the developmental pathway? Most of the data included in the CHILDES database come from one hour recordings that were prepared at regular intervals of one to three weeks. Using this sampling method, the transcripts represent between 1.0% to 1.5% of the language a child hears and produces during the time of the study (cf. Tomasello and Stahl 2004). While 1.5% of the data may provide enough information to trace the development of frequently occurring structures and expressions, they are not sufficient to
study the development of less frequent phenomena. As Tomasello and Stahl (2004) have shown, in order to trace the development of infrequent grammatical patterns one needs to examine a much larger proportion of the child’s speech. In particular, the age of appearance and the order of acquisition are not reliable if we use the standard language samples of one or two hours recorded every other week; rather, what we need are much denser data collections, as they are currently prepared at the Max Planck Institute for Evolutionary Anthropology in Leipzig.

2. Sample studies

2.1. The acquisition of complex sentences

While observational studies have a long tradition in child language research, the acquisition of many grammatical phenomena has been primarily studied based on data from experiments. A case in point is the acquisition of complex sentences. Complex sentences are grammatical constructions consisting of a matrix clause and a subordinate clause, i.e. a relative clause, a complement clause, or an adverbial clause. While the acquisition of all of these constructions has been studied extensively in various experiments, the developments of relative and complement clauses have never been investigated based on observational data until recently. Using the resources of the CHILDES database, Diessel (2004) provides the first comprehensive
study to systematically analyze the acquisition of all complex sentences in spontaneous child language. Interestingly, he found that most subordinate clauses are embedded in particular main clause constructions. For instance, the earliest relative clauses children produce are attached to the predicate nominal of a copular clause, or, less frequently, to an isolated noun phrase (cf. examples 1a-b).

(1)  

a. The pig bumps into the horse that jumps over the giraffe.

b. The lion that jumps over the pig bumps into the lion.

Although the sentences in (1a-b) consist of two clauses, they only contain a single proposition. The copular clause does not denote an independent situation; rather, it functions to establish a referent in focus position making it available for the predication expressed in the relative clause. These relative constructions are much simpler than the ones that have been used in most experiments, in which relative clauses are commonly attached to the subject or object of a full-fledged main clause (e.g. The duck that kissed the rabbit jumped over the fence). This explains why even four an five year-old children have great difficulties in comprehending relative constructions in experiments despite the fact that they begin to use them quite early. As Diessel and Tomasello (2004) have shown in a recent experimental study, when children are confronted with relative constructions similar to the ones
that they produce in spontaneous speech, their performance improves significantly compared to their performance in previous experimental studies. Thus, while experiments play a key role in the study of child language, they should always be combined with thorough corpus-based studies.

2.2. The ambient language

Another topic that has been re-evaluated based on recent corpus-based analyses is the role of the ambient language. According to Chomsky (1972), there is an enormous discrepancy between the ambient language and the grammar children learn. Although Chomsky never studied the ambient language, he claimed that the data children receive is not sufficient to learn grammar from experience alone. Based on this hypothesis, he suggested that the ‘gap’ between the insufficient input data and the target grammar can only be closed because children are endowed with an innate linguistic grammar. This hypothesis, which has become known as the ‘poverty of the stimulus’, is perhaps the most important argument for the nativist approach to language acquisition (see Diessel 2004: chap 2 for a recent review).

While Chomsky’s dictum had an enormous effect on the study of language acquisition, recent corpus-based analyses suggest that the ambient language provides more systematic information than commonly assumed. Investigating some of the constructions that according to Chomsky and other
nativists are so infrequent that they cannot be learned from experience, Pullum and Scholz (2002) showed that all of these constructions are well-attested in adult language. Since Pullum and Scholz did not specifically investigate the ambient language, their data do not immediately bear on the argument from the poverty of the stimulus. However, other recent investigations have shown that the ambient language provides rich distributional cues for many grammatical phenomena. For instance, Mintz, Newport and Bever (2002) have shown that there are striking distributional differences between nouns and verbs in the ambient language. If we assume that children are able to recognize the distributional regularities provided by the ambient language, they should be able to establish the major word classes based on distributional learning. Similar hypotheses have been suggested in other recent corpus-based analyses for other grammatical phenomena (cf. Brent and Cartwright 1996; Cartwright and Brent 1997; Reddington and Chater 1998; Raddington, Chater and Finch 1998).
References


This chapter provides an overview of corpus-linguistic approaches to the study of language acquisition.

to assess a particular aspect of children’s linguistic knowledge, or they use observational data to

In order to assess the development of children’s linguistic knowledge, researchers use a variety of methods. but most investigations begin with the inspection of observational data (cf. McDaniel et al. 1996). If these data are transcribed and compiled in machine-readable from they can be seen as a linguistic corpus, but it must be emphasized that they are different from corpora of adult corpus linguistics.

This can be longitudinal data of children that were studied over an extended period of time, or it can be cross-sectional data from children of different age groups that were all collected at the same time.

Young children use language primarily in conversations with adult speakers, notably their parents, and this is the kind of data we usually find in a child language corpus.
This paper provides an overview of corpus-based approaches to the study of language acquisition. It is divided into three parts. The first part describes the particular properties of child language data and the criteria that are important to build a representative child language corpus. The second part describes the CHILDES database, which is the most important corpus of child language. And the third part presents two recent sample studies that have drawn on data from the CHILDES database.

However, such a child language corpus is very different from a corpus of adult language.