PhonBank Manual
A Guide to the PhonBank Database

Web access:
PhonBank CHAT data: http://childes.psy.cmu.edu/data/PhonBank/
PhonBank Phon data: http://childes.psy.cmu.edu/data/PhonBank-Phon/
PhonBank media: http://childes.psy.cmu.edu/media/media/PhonBank/
CLAN Application: http://childes.psy.cmu.edu/clan/
Phon Application: http://phon.ling.mun.ca/phontrac/wiki/Downloads

Document prepared by Carla Peddle
In collaboration with Brian MacWhinney & Yvan Rose

Last revised: August 25, 2014
# Table of Contents

Dutch-Levelt & Fikkert .................................................................................. 3
Dutch-Zink .................................................................................................. 4
English-Chiat .............................................................................................. 5
English-Compton & Streeter .................................................................... 6
English-Davis ............................................................................................. 7
English-Inkelas ......................................................................................... 10
English-Smith ......................................................................................... 11
English-Stanford ..................................................................................... 12
French-Goad & Rose ................................................................................. 13
French-Kern .............................................................................................. 14
French-Stanford ....................................................................................... 15
German-PAIDUS ..................................................................................... 16
German-Stuttgart/TAKI ............................................................................. 18
Japanese-Ota ............................................................................................ 21
Japanese-Stanford .................................................................................. 22
PAIDOLOGOS-Cantonese/English/Greek/Japanese .................................... 23
Portuguese-Lisbon ................................................................................... 24
Portuguese-Freitas .................................................................................. 26
Romanian-Kern ....................................................................................... 27
Spanish-PAIDUS ..................................................................................... 28
Swedish-Stanford .................................................................................... 30
Tunisian-Kern ......................................................................................... 32
Dutch-Levelt & Fikkert

**Project Name:** Dutch-CLPF

**Investigators:** Claartje Levelt    Paula Fikkert

**Contact:** C.C.Levelt@hum.leidenuniv.nl    paula.fikkert@mac.com

**Location:** Leiden & Groningen, NL

**Number of Participants:** 12

**Nature of the Study:** babbling, full word, naturalistic, longitudinal

**Media Type (if any):** wave files

**Citation Information:**


<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catootje</td>
<td>1988-10-24</td>
<td>1;10.10 - 2;07.04</td>
<td>16</td>
<td>F</td>
</tr>
<tr>
<td>David</td>
<td>1987-11-23</td>
<td>1;11.07 - 2;03.27</td>
<td>6</td>
<td>M</td>
</tr>
<tr>
<td>Elke</td>
<td>1988-04-06</td>
<td>1;06.24 - 2;04.28</td>
<td>19</td>
<td>F</td>
</tr>
<tr>
<td>Enzo</td>
<td>1987-11-30</td>
<td>1;11.08 - 2;06.11</td>
<td>16</td>
<td>M</td>
</tr>
<tr>
<td>Eva</td>
<td>1988-06-06</td>
<td>1;04.12 - 1;11.08</td>
<td>12</td>
<td>F</td>
</tr>
<tr>
<td>Jarmo</td>
<td>1988-05-03</td>
<td>1;04.18 - 2;04.01</td>
<td>23</td>
<td>M</td>
</tr>
<tr>
<td>Leon</td>
<td>1987-12-10</td>
<td>1;10.01 - 2;08.19</td>
<td>23</td>
<td>M</td>
</tr>
<tr>
<td>Leonie</td>
<td>1986-05-07</td>
<td>1;09.15 - 1;11.18</td>
<td>7</td>
<td>F</td>
</tr>
<tr>
<td>Noortje</td>
<td>1988-02-10</td>
<td>1;07.14 - 2;11.00</td>
<td>21</td>
<td>F</td>
</tr>
<tr>
<td>Robin</td>
<td>1988-05-27</td>
<td>1;05.10 - 2;04.29</td>
<td>23</td>
<td>F</td>
</tr>
<tr>
<td>Tirza</td>
<td>1988-04-13</td>
<td>1;07.09 - 2;06.12</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>Tom</td>
<td>1988-10-05</td>
<td>1;00.24 - 2;03.02</td>
<td>25</td>
<td>M</td>
</tr>
</tbody>
</table>

Recordings were made in the child’s home during natural, spontaneous, interactive sessions with one or both of the experimenters and occasionally with one of the parents. Typically a researcher or parent would interact with the child by reading books or playing with toys and occasionally asking the child what s/he saw in books or what s/he was doing.

During transcription, unintelligible utterances were left out, however, false starts, errors, breakdowns, etc. were transcribed since they may provide valuable information about the child’s competence. If the child’s utterances contained more than one word, boundaries were placed between the words. No boundaries were placed if it was unclear whether the child knew the different words of which the utterance consisted, such as “what is that?”.

This corpus was converted from the original CLPF database. We extend many thanks to Claartje Levelt and Paula Fikkert for their assistance with the conversion process. The transcripts have been linked to the available audio clips for five children: Catootje, Enzo, Eva, Robin, and Tirza. We hope to add the audio for the other children at a later date.
At this stage, the data have not been entirely rechecked from the original transcriptions. Careful spot-checking however suggests that the conversion is accurate. Please feel free to contact us or the original authors in case you have questions.

The following lists illustrate some child-specific forms and onomatopoeia present in the childrens’ speech.

<table>
<thead>
<tr>
<th>Child-specific forms</th>
<th>Referent</th>
<th>Onomatopoeia</th>
<th>Referent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertje</td>
<td>piglet</td>
<td>mei</td>
<td>Sheep, goat</td>
</tr>
<tr>
<td>Bumba</td>
<td>clown</td>
<td>pok</td>
<td>chicken</td>
</tr>
<tr>
<td>Dribbel</td>
<td>dog</td>
<td>hi</td>
<td>horse</td>
</tr>
<tr>
<td>Jules</td>
<td>boy</td>
<td>kwak</td>
<td>duck</td>
</tr>
<tr>
<td>Willy</td>
<td>animal</td>
<td>ia</td>
<td>donkey</td>
</tr>
<tr>
<td>Zina</td>
<td>dog</td>
<td>huuu</td>
<td>horse</td>
</tr>
<tr>
<td>Lola</td>
<td>chicken</td>
<td>woef</td>
<td>dog</td>
</tr>
<tr>
<td>Tiny</td>
<td>girl</td>
<td>miauw</td>
<td>cat</td>
</tr>
<tr>
<td>Bambi</td>
<td>deer</td>
<td>kukeleku</td>
<td>rooster</td>
</tr>
<tr>
<td>Musti</td>
<td>cat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sofie</td>
<td>crocodile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samson</td>
<td>dog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sjoewie</td>
<td>dog</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

---

**Dutch-Zink**

**Project Name:** Dutch-Zink

**Investigators:** Inge Zink

**Contact:** inge.zink@med.kuleuven.be

**Location:** Leuven, Brabant and Antwerp, BE

**Number of Participants:** 4

**Nature of the Study:** naturalistic, longitudinal

**Media Type (if any):** Mini DV Cassettes (later digitized to DVDs)

**Citation Information:**


<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>2002-09-03</td>
<td>08.04 – 2;02.14</td>
<td>38</td>
<td>M</td>
</tr>
<tr>
<td>Judith</td>
<td>2003-06-13</td>
<td>08.03 – 2;00.26</td>
<td>39</td>
<td>F</td>
</tr>
<tr>
<td>Laurien</td>
<td>2002-03-08</td>
<td>08.07 – 2;01.26</td>
<td>37</td>
<td>F</td>
</tr>
<tr>
<td>Meinder</td>
<td>2002-05-07</td>
<td>08.14 – 2;02.05</td>
<td>38</td>
<td>M</td>
</tr>
</tbody>
</table>

The recordings for this corpus were made in Leuven, Brabant, Belgium (3 children: Meinder, Judith, Laurien) and Antwerp, Belgium (1 child: David). The participants were recorded every two weeks from 8 months to 25 months of age. Each recording session lasted approximately 60 minutes. The team made recordings using a Panasonic Digital Video Camera (NV-GS180) and a Sennheiser ew100 microphone on 60 minute fuji Mini DV Cassettes. The video files were later digitized to DVDs. During the recording sessions, the child interacted with his/her mother (or both parents) and a speech therapist and/or a master's student of speech-language pathology was present.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

**English-Chiat**

**Project Name:** English-Chiat

**Investigator(s):** Shula Chiat

**Contact:** shula.chiat.1@city.ac.uk

**Location:** UK

**Number of Participants:** 3

**Nature of study:** clinical, naturalistic, longitudinal

**Media Type (if any):** Audio files (wave files)

**Citation Information:**
Recordings for this project were carried out in a quiet room at school or a clinic. Each child can be considered its own single case study. Utterances were spontaneous in nature through play sessions with the researcher. There were also some elicited repetitions of real words, non-words and phrases to sample the realization of segmental targets in different prosodic positions at the word- and phrase-level.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>DI</td>
<td>1983-04-06</td>
<td>5;05.16 – 5;05.28</td>
<td>4</td>
<td>M</td>
</tr>
<tr>
<td>SB</td>
<td>1982-11-27</td>
<td>5;05-26 – 5;07.09</td>
<td>4</td>
<td>M</td>
</tr>
<tr>
<td>SR</td>
<td>1976-01-30</td>
<td>5;07.21 – 5;08.12</td>
<td>3</td>
<td>M</td>
</tr>
</tbody>
</table>

A team, under the direction of A. J. Compton, in the 1970s originally collected these data. The method of data collection and some preliminary analyses, were presented in Compton & Streeter
The project was undertaken to map out, as precisely as possible, the development of children’s sound systems. With this goal in mind, a diary method of data collection was chosen, with parents keep track of their children’s utterances by recording them in notebooks at least four days a week and scattered throughout the child’s waking hours, covering about four hours a day. The parents were speech pathologists and received additional training in the phonetic transcription of child speech prior to the study.

All of the children were learning American English as spoken in California. None of the children had any language or learning-related impairments.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

---

**English-Davis**

**Project Name:** English-Davis  
**Investigator(s):** Barbara Davis  
**Contact:** babs@mail.utexas.edu  
**Location:** Austin, TX  
**Number of Participants:** 21  
**Nature of study:** naturalistic, longitudinal  
**Media Type (if any):** Reel-to-reel and DAT Audio tape

**Citation Information:**  
MacNeilage, P.F. & Davis, B.L. (2001). Commentary: The role of rhythmic cyclicities in infant


<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaron</td>
<td>1981-01-14</td>
<td>1;08.01 – 2;01.03</td>
<td>11</td>
<td>M</td>
</tr>
<tr>
<td>Anthony</td>
<td>1981-05-03</td>
<td>1;08.05 – 2;01.15</td>
<td>13</td>
<td>M</td>
</tr>
<tr>
<td>Ben</td>
<td>2002-04-06</td>
<td>0;11.21 – 2;04.02</td>
<td>21</td>
<td>M</td>
</tr>
<tr>
<td>Cameron</td>
<td>1992-02-15</td>
<td>0;07.11 – 2;11.24</td>
<td>52</td>
<td>F</td>
</tr>
<tr>
<td>Charlotte</td>
<td>2002-10-29</td>
<td>0;10.12 – 2;11.22</td>
<td>44</td>
<td>F</td>
</tr>
<tr>
<td>Georgia</td>
<td>2002-08-27</td>
<td>0;08.25 – 2;11.05</td>
<td>45</td>
<td>F</td>
</tr>
<tr>
<td>Hannah</td>
<td>2002-08-23</td>
<td>0;11.14 – 2;04.24</td>
<td>29</td>
<td>F</td>
</tr>
<tr>
<td>Jodi</td>
<td>1981-09-07</td>
<td>1;00.29 – 1;07.09</td>
<td>14</td>
<td>F</td>
</tr>
<tr>
<td>Kaeley</td>
<td>2002-11-21</td>
<td>1;00.24 – 2;01.23</td>
<td>13</td>
<td>F</td>
</tr>
<tr>
<td>Kate</td>
<td>1981-05-29</td>
<td>1;03.02 – 1;07.25</td>
<td>10</td>
<td>F</td>
</tr>
<tr>
<td>Martin</td>
<td>1993-02-22</td>
<td>1;05.19 – 2;02.09</td>
<td>13</td>
<td>M</td>
</tr>
<tr>
<td>Micah</td>
<td>1988-09-13</td>
<td>0;08.01 – 1;06.19</td>
<td>29</td>
<td>M</td>
</tr>
<tr>
<td>Nate</td>
<td>2001-08-06</td>
<td>0;10.07 – 2;09.07</td>
<td>26</td>
<td>M</td>
</tr>
<tr>
<td>Nick</td>
<td>1991-12-09</td>
<td>0;10.20 – 3;01.02</td>
<td>46</td>
<td>M</td>
</tr>
<tr>
<td>Paxton</td>
<td>1992-03-19</td>
<td>0;08.02 – 2;00.02</td>
<td>47</td>
<td>M</td>
</tr>
<tr>
<td>Rachel</td>
<td>1992-01-30</td>
<td>0;08.04 – 1;01.02</td>
<td>40</td>
<td>F</td>
</tr>
<tr>
<td>Rebecca</td>
<td>1983-09-17</td>
<td>1;01.24 – 1;07.17</td>
<td>41</td>
<td>F</td>
</tr>
<tr>
<td>Rowan</td>
<td>2001-11-11</td>
<td>0;10.23 – 2;10.19</td>
<td>41</td>
<td>M</td>
</tr>
<tr>
<td>Sadie</td>
<td>1992-04-14</td>
<td>0;07.11 – 1;07.03</td>
<td>33</td>
<td>F</td>
</tr>
<tr>
<td>Sam</td>
<td>2001-08-24</td>
<td>0;10.07 – 2;01.03</td>
<td>26</td>
<td>M</td>
</tr>
<tr>
<td>Will</td>
<td>1992-03-27</td>
<td>0;07.29 – 1;04.13</td>
<td>24</td>
<td>M</td>
</tr>
</tbody>
</table>

Data analyzed for this study were collected as a part of a longitudinal project tracking early normal speech development from the onset of canonical babbling through age 3.5 years. Subjects were located by informal referral from the surrounding community. Normal development was established through parent case history report. In addition, each infant was administered the Battelle Developmental Screening Inventory (Guidubaldi, Newborg, Stock, Svinicki, & Wneck, 1984) and hearing screening using sound field techniques.

Video and audio recordings were gathered in the children’s homes during natural interactions and situations that occurred in their daily lives. The children were interacting with their parent or others present in their home setting. The experimenter sometimes interacted with the child and parent and sometimes observed as seemed natural in the setting.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.
English-Inkelas

Project Name: English-Inkelas
Investigator(s): Sharon Inkelas
Contact: inkelas@berkeley.edu
Location: Northern California
Number of Participants: 1
Nature of study: naturalistic, diary, longitudinal
Media Type (if any): none

Citation Information:

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eli</td>
<td>1997-12-22</td>
<td>0;06.08 - 3;09.28</td>
<td>200</td>
<td>M</td>
</tr>
</tbody>
</table>

This study is based primarily on a new longitudinal corpus from E, a typically developing monolingual learner of English with no history of hearing problems or language delays. E has one sibling, a brother who is two years and five months older and who also exhibited normal phonological development. E’s mother is a native speaker of American English. E’s father is a native speaker of Turkish, but fluent in English; his slight Turkish accent is unlikely to have been a factor in E’s phonological patterns. During the course of language development, E was raised in a monolingual English-speaking environment. He was exposed to minimal amounts of Turkish but was virtually never addressed directly in this language. As of age 2;7, due to his sibling’s entry into a Spanish immersion school program, E began to be exposed to some Spanish, including the word *hola*, but not by native Spanish speakers.

The data from E were gathered in a naturalistic, diary setting primarily by his mother, a trained phonologist. E’s father, also a phonologist, participated in the data collection as well. Data collection was performed as often as possible, during unstructured sessions and regular family activities. The period of data gathering spans E’s ages 0;06.09 to 3;09.29, with most of the data collected prior to the last year of the study. Between ages 0;06.09 and 2;09.09, a span that corresponds to E’s phonological development peak (i.e. during which most aspects of the target grammar were acquired), a total of 3,267 words (distributed over 1,713 utterances) were transcribed phonetically. An additional 113 words were transcribed over the last year of the study, during which E’s mother was attending mainly to the development of /l/.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

_____________________________________________________________________________
The child in the case study, Amahl, is the son of an English father with some derivations from “received pronunciations” and Indian mother who speaks English only as a fourth language (following Hindi, Bengali, Marathi). Amahl was born in Boston, Massachusetts and moved to the U.K. at approximately one year of age, which may have influenced his late speaking. The child lived with his paternal aunt from 19 months to 2 years, and started attending day nurseries and play groups in Bedfordshire and Hertfordshire at 30 months.

The data discussed were recorded in IPA transcriptions on index cards. Some tape recordings were made occasionally, however, most of the description is based on non-recorded material. There were instances of spontaneous speech, however the majority of the words were elicited with phrases such as “can you say “zink” for me?”.

This corpus was retyped and adapted from the original diary study. A significant portion of this work was undertaken by Tania Zamuner; additional work was performed under an FCAR grant to Heather Goad. The recording dates in the corpus correspond to the first date of Smith’s (1973) descriptive stages. Please note that the data have not been entirely rechecked from Smith’s original transcriptions. Careful spot-checking however suggests that the conversion is accurate. Please feel free to contact us in case you have questions or if you notice any errors.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by the above reference.
English-Stanford

Incomplete

Project Name: English-Stanford
Investigator(s): Marilyn Vihman
Contact: mv509@york.ac.uk
Location: Stanford area, CA
Number of Participants: 5
Nature of study: naturalistic, longitudinal
Media Type (if any): audio cassettes and/or reel-to-reel/VHS video

Citation Information:

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deb</td>
<td>TBA</td>
<td>0:09.17 – 1:03.24</td>
<td>6</td>
<td>F</td>
</tr>
<tr>
<td>Emi</td>
<td>TBA</td>
<td>0:10.07 – 1:02.39</td>
<td>6</td>
<td>F</td>
</tr>
<tr>
<td>Mol</td>
<td>TBA</td>
<td>0:08.17 – 1:02.20</td>
<td>6</td>
<td>F</td>
</tr>
<tr>
<td>Sea</td>
<td>TBA</td>
<td>0:10.00 – 1:03.23</td>
<td>6</td>
<td>M</td>
</tr>
<tr>
<td>Tim</td>
<td>TBA</td>
<td>0:10.08 – 1:04.22</td>
<td>6</td>
<td>M</td>
</tr>
</tbody>
</table>

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.
French-Goad & Rose

Project Name: French-GoadRose

Investigators: Heather Goad  Yvan Rose

Contact: heather.goad@mcgill.ca  yrose@mun.ca

Location: Québec City & just outside Montréal, QC

Number of Participants: 2

Nature of the Study: babbling to full word, naturalistic, longitudinal

Media Type (if any): wave files

Citation Information:

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clara</td>
<td>1996-03-26</td>
<td>1;00.27-2;07.19</td>
<td>34</td>
<td>F</td>
</tr>
<tr>
<td>Théo</td>
<td>1994-11-25</td>
<td>1;10.26-4;00.00</td>
<td>45</td>
<td>M</td>
</tr>
</tbody>
</table>

The participants in this project were recorded in their homes in a naturalistic setting, generally in the absence of siblings, approximately every second week. Clara was recorded by her mother, a sociolinguist and Théo was recorded by a female French-speaking family member who was a linguistics graduate student from McGill.

During the recordings the children typically looked at picture books and played with toys. The interviewer encouraged spontaneous word productions by the child to avoid speech sample repetition, and repeated the words produced by the child to facilitate identification of data for extraction and transcription.

The recordings were made using an analog recording machine and a multidirectional microphone, which was set up near the child on a foamy cushion the floor to reduce interfering noises from movements and toys. The tapes were digitized using SoundEdit 16v2 in 16 bit sample size at a rate of 22050 kHz.

This corpus was converted from the original database of Québec French used by Rose (2000). The data were collected under an FCAR grant to Heather Goad. Please note that the data have not been entirely rechecked from the original transcriptions. Careful spot-checking however suggests that the conversion is accurate. Please feel free to contact the contributor in case you have questions.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by the above reference.
French-Kern

Project Name: French-Kern
Investigator(s): Sophie Kern
Contact: Sophie.Kern@univ-lyon2.fr
Location: Lyon, FR
Number of Participants: 4
Nature of study: naturalistic, longitudinal, monolingual
Media Type (if any): digital video (.mov, .mp4)

Citation Information:

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baptiste</td>
<td>2000-09-11</td>
<td>0:09.07 - 02:00.12</td>
<td>35</td>
<td>M</td>
</tr>
<tr>
<td>Emma</td>
<td>2000-10-12</td>
<td>0:08.16 - 2:01.08</td>
<td>36</td>
<td>F</td>
</tr>
<tr>
<td>Esteban</td>
<td>2001-10-23</td>
<td>0:07.15 – 2:00.24</td>
<td>30</td>
<td>M</td>
</tr>
<tr>
<td>Jules</td>
<td>2001-03-28</td>
<td>0:09.13 - 2:00.28</td>
<td>28</td>
<td>M</td>
</tr>
</tbody>
</table>

Five types of data were collected. First, one hour of spontaneous vocalization data was audio and video recorded every two weeks from 8 months of age through 25 months of age. Recording took place in children’s homes. The parents were told to follow their normal types of activities with their child. Second, minimally 1,000 dictionary entries from the ambient language employed by the parents of each child participant were analyzed for comparison with the child data for that language. Parental reports were administered using adaptations of the MacArthur Development Inventories (Fenson et al., 1993) respectively elaborated for Dutch for French participants. Mothers filled out the questionnaire once in a month. For the remaining languages there is no adaptation yet, but one could imagine using the spontaneous data to elaborate the same instrument. An object manipulation categorization task was administered every two months. This task was conceived to evaluate the children’s spontaneous nonverbal categorization abilities. Several toys, which were consistent across the language groups served as stimuli. Each task involved a contrast of objects from two different categories (animal, means of locomotion, furniture).

Children were developing normally by community standards as well as reports from parents and physicians regarding developmental milestones were observed in the normal daily environment.
These children were becoming monolingual speakers of French, Romanian and Tunisian Arabic. Languages were chosen to include different language families as well as for the contrasts they show in phonetic and phonological features of interest for early language development. These characteristics include word length, syllabic types as well as phonemic inventory diversity.

One hour of spontaneous vocalization data was audio and video recorded every two weeks from 8 months of age through 25 months of age. Recording took place in children's homes. The parents were told to follow their normal types of activities with their child. After collection, data was phonetically transcribed using the International Phonetic Alphabet (IPA). Broad phonetic transcriptions were used, supplemented by some diacritics (mainly for palatalized, pharyngealized, nasalized sounds and duration of sounds). Tokens considered as single utterance strings were bounded by one second of silence, noise or adult speech.

This work was supported by the EUROCORE Program “The Origin of Man, Language and Languages” (OMLL) and the French CNRS program “Origine de l’Homme, du Langage et des Langues” (OHLL).

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

---

**French-Stanford**  
**Incomplete**

**Project Name:** French-Stanford  
**Investigator(s):** Marilyn Vihman  
**Contact:** mv509@york.ac.uk  
**Location:** Paris, FR  
**Number of Participants:** 5 (or 6)  
**Nature of study:** naturalistic, longitudinal  
**Media Type (if any):** audio cassettes and/or reel-to-reel/VHS video

**Citation Information:**  

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
</table>
In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

**German-PAIDUS**

**Project Name:** German-PAIDUS

**Investigator(s):**
- Conxita Lleó
- Antonio Maldonado

**Contact:**
- lleo@uni-hamburg.de
- Universidad Autónoma de Madrid

**Location:** Hamburg, Germany

**Number of Participants:** 5

**Nature of study:** naturalistic, babbling, first words, spontaneous, informal testing

**Media Type (if any):** Wave files

**Citation Information:**


Lleó, Conxita (2006). The Acquisition of Prosodic Word Structures in Spanish by Monolingual
and Spanish-German Bilingual Children. *Language and Speech* 49 (2), 207-231.


<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernd</td>
<td>1989-12-15</td>
<td>0:08.22 – 3:02.24</td>
<td>34</td>
<td>M</td>
</tr>
<tr>
<td>Britta</td>
<td>1989-08-10</td>
<td>0:09.01 – 3:03.15</td>
<td>38</td>
<td>F</td>
</tr>
<tr>
<td>Johannes</td>
<td>1989-10-31</td>
<td>0:09.00 – 2:10.10</td>
<td>30</td>
<td>M</td>
</tr>
<tr>
<td>Marion</td>
<td>1989-08-14</td>
<td>0:08.27 – 3:03.11</td>
<td>35</td>
<td>F</td>
</tr>
<tr>
<td>Thomas</td>
<td>1989-08-06</td>
<td>0:09.26 – 3:05.12</td>
<td>38</td>
<td>M</td>
</tr>
</tbody>
</table>

The project BIDS/PAIDUS was carried out at the University of Hamburg (Department of Romance Languages) with the support of the Deutsche Forschungsgemeinschaft under the direction of Conxita Lleó. The project began in 1990 first as BIDS, with the goal of comparing the phonetic characteristics of babbling in German and comparing it with babbling in Spanish. With this purpose, a collaboration with Antonio Maldonado of the Universidad Autónoma de Madrid (Dept. of Psychology) was started and we applied for a joint grant from “Acciones Integradas”, so that the two teams could visit each other. The German data were collected in Hamburg, Germany, and the Spanish data in Madrid, Spain.

All five German children had been born in Hamburg and lived there. Their parents were all German and spoke the Northern variety of German. The children were visited from very early on, before they began to produce their first words, as we wanted to study their babbling. Some informal tests, in order to figure their (passive) vocabularies were conducted. After the children began saying words, we continued recording them within the project PAIDUS, which had as its goal to finding out how phonological parameters were possibly fixed for German as opposed to Spanish. The German children were visited every 2 weeks at first, and later on once a month. Recordings were done at the children’s homes by two project research assistants. At first, the mother used to be present, but later on, we made most of the recordings with the child alone interacting with one research assistant. We tried to record the child under spontaneous conditions, playing, chatting, etc. We also brought some objects (little animals or toys) in order to elicit certain words from all children, making their phonetic and phonological characteristics comparable. The mother tried to keep a record of possible new words that were new in the child’s vocabularies, and we tried to elicit them, as well. We also used books with pictures, which belonged to the child and others that we brought. The recordings were audio recordings, and in order not to miss the ambient setting and context, one of the researchers was writing down all possible details, including what the child was doing, while the other assistant was interacting with the child.

Both projects, BIDS and PAIDUS, were financed by the DFG with grant to Conxita Lleó. BIDS from 1989 until 1991 and PAIDUS from 1991 until 1994. Research Assistants in the project BIDS were Christliebe El Mogharbel and Kerstin Feuge. Research Assistants in the project PAIDUS were Christliebe El Mogharbel and Michael Prinz.
Researchers using the BIDS/PAIDUS data should send a copy of their publication based on the
data to the contributor of the data. In accordance with CHILDES rules, any use of data from this
corpus must be accompanied by at least one of the above references.

---

**German-Stuttgart/TAKI**

**Project Name:** German-Stuttgart/TAKI  
**Investigator(s):** Bernd Möbius  
**Contact:** bernd.moebius@ims.uni-stuttgart.de  
**Location:** Stuttgart area, Southern Germany  
**Number of Participants:** 11  
**Nature of study:** naturalistic, longitudinal, spontaneous, parental interaction, naming  
**Media Type (if any):** DAT tapes  

**Citation Information:**

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maarten (MS)</td>
<td>1998-09-08</td>
<td>3;9 – 8;0</td>
<td>40</td>
<td>M</td>
</tr>
<tr>
<td>Lena (LL)</td>
<td>1999-03-23</td>
<td>3;3 – 7;6</td>
<td>46</td>
<td>F</td>
</tr>
<tr>
<td>Tjark (TS)</td>
<td>2000-11-19</td>
<td>1;6 – 2;4</td>
<td>42</td>
<td>M</td>
</tr>
<tr>
<td>Evelyn (ED)</td>
<td>2002-02-21</td>
<td>2;3 – 4;7</td>
<td>105</td>
<td>F</td>
</tr>
<tr>
<td>Robin (RL)</td>
<td>2002-02-21</td>
<td>3;2 – 4;7</td>
<td>58</td>
<td>M</td>
</tr>
<tr>
<td>Emma (EL)</td>
<td>2002-06-22</td>
<td>0;5 – 4;3</td>
<td>7</td>
<td>F</td>
</tr>
<tr>
<td>Hannah (HH)</td>
<td>2003-08-02</td>
<td>0;4 – 3;1</td>
<td>11</td>
<td>F</td>
</tr>
<tr>
<td>Olli (OZ)</td>
<td>2004-03-01</td>
<td>0;8 – 2;9</td>
<td>10</td>
<td>M</td>
</tr>
<tr>
<td>Bennie (BW)</td>
<td>2004-04-22</td>
<td>0;11 – 2;11</td>
<td>32</td>
<td>M</td>
</tr>
<tr>
<td>Nils (NB)</td>
<td>2004-08-20</td>
<td>0;8 – 2;8</td>
<td>32</td>
<td>M</td>
</tr>
<tr>
<td>Rike (FZ)</td>
<td>2005-09-05</td>
<td>1;0 – 2;6</td>
<td>14</td>
<td>F</td>
</tr>
</tbody>
</table>

The Stuttgart Corpus was constructed as part of a study on the acquisition of stress in German
founded by the German Research Foundation. In this project the researchers recorded and
analyzed acoustic data of 11 children from 6 months up to 15 years to develop an exemplar-
based model of stress acquisition in German and to build up a prosodic annotated speech corpus
of babbling, first words and meaningful speech from German speaking children. The following
table summarizes information about the children and the recordings.
The recordings were made at the children's homes in familiar play situations with their parents. We recorded using a Sony DAT TCD-D100 and a high-quality wireless microphone NADY LT-4 (Lavalier) E-701 (600 Ohm). We tried to keep the distance between the microphone and the mouth as constant as possible during the entire session to obtain high-quality recordings for acoustic analysis. The recorded data were transferred to a computer and down sampled to 16 kHz.

1. Babbling and first words: 5–18 months of age
Between 5 and 18 months of age, the speech data from six children (3 boys, 3 girls) were collected. The infants were audio-recorded every 6–8 weeks, starting between five and seven months, when first CV-syllable productions occur, at their homes in familiar play situations with their parents. No person unfamiliar to the child was present during the recordings. During a recording session a parent (normally the mother) played with the child and later on, at the one-word stage, looked and talked about a picture book and picture cards. Before the recordings started, the parents were instructed on how to use the recording equipment. The parents' utterances were recorded and analyzed as well. All of the infants lived in monolingual German-speaking families and had no unusual prenatal, sensory or developmental concerns or hearing problems. At the age of 12 months, the speech development of all infants was tested using a parental questionnaire for early recognition of children at risk (Grimm & Doil, 2004).

2. Mixing-phase: 18–36 months of age
Between 18 and 36 months of age, we collected speech data every 6–8 weeks. Because we have different recording tasks developed for this age we called this phase mixing. The recordings were done until the third birthday in the way described for the babbling phase. The children born in 2004 (OZ, BW, NB, FZ) were also tested at each recording session for their use of stress in multisyllabic words.

3. Card Naming (from 18 months of age)
We created picture cards representing two- and three-syllable German words with stress on the first, second and third syllable with the vowels /a/, /i/, and /o/ in stressed and unstressed position. From the age of 18 months, children were tested with these cards. With this task, the development of different stress schemes can be tested and described. The words were: Akrobat, Anorak, Banane, Bikini, Buchstabe, Eisenbahn, Elefant, Eskimo, Fliege, Flughafen, Fotograf, Gardine, Giraffe, Gitarre, Gorilla, Kamel, Kanone, Kobolde, Kokosnuss, Korkodil, Lastwagen, Lawine, Malerin, Matrose, Mikrofon, Müllwagen, Paket, Papagei, Pinguin, Pistole, Polizei, Postauto, Postbote, Prinzessin, Pullover, Radfahrer, Rennfahrer, Sandale, Saxophon, Schmetterling, Skifahrer, Spiegel, Stadion, Teddybär, Tomate, Trampolin, Trompete, Vulkan(e), Wohnwagen, and Zitrone.
4. TAKI task: from 36 months of age

Beginning at 36 months, recordings were made each 10–12 weeks using the TAKI task proposed by Allen (1980). We created five pairs of animal toys and assigned nonce names to each. Within each pair, the nonce names differed only in terms of the position of main stress. For example, in a pair with a brown bear and a polar bear. Both were called “bimo”, but the stress was on the first syllable for the brown bear and the second syllable for the polar bear. The nonce names were all bi-syllabic or tri-syllabic and consisted of consonant-vowel (CV) syllables formed from vowels /a i o/ and consonants /b d m n/. All of these stress patterns are possible in German, although stress on the second or third syllable is much less common than stress on the first syllable.

Annotation labels
1. CV coding (.cv). To study the development of syllable structure, we marked the beginning and ending of each vowel and consonant.
2. Stable Vowel (.marks). The beginning and end of the stable phase of the vowel is marked. No influence of the surrounding context should be hearable. The stable phase of a vowel is characterized by parallel formants (observable in the spectrogram) and a constant waveform (observable in the time signal). VA (Anfang) marks the beginning of the vowel and VE (Ende) the ending. The indexes “u” are used for unstressed (unbetont), “b” for stressed (betont) vowels.
3. Stress (.stress): Perceptual prominence for each syllable: no prominence (0), most prominent (1), prominent but not most (2).
4. Orthographic transcription (.trans). Each utterance was annotated on the syllabic level by two trained transcribers.
5. Cover Symbol transcription (.phones-cover). Manner and place of articulation describes the consonantal structure: L=labial, A=Alveolar, V=velar, G=glottal, O=other; P=plosive, N=nasal, F=fricative, G=glide O=other. Tongue height and manner describes the vocalic structure: H=high, M=medium, T=low, tief. V=front, vone, Z=central, zentrale, H=back, ?
6. Narrow transcription in XSAMPA (.phones).
7. Syllable structure coding (.sylstr): Syllable structure, words and syllable marker.
   - Aw = beginning of utterance (Anfang)
   - Ew = end of utterance (Ende)
   - S = syllable border
   - 1 = precanonical babbling
   - 2 = reduplicated babbling, CV-syllable with real consonants ([mamma], [dae])
   - 3 = one or more CV-syllables with real consonant with different articulation place and/or manner ([min], [dajaep])
   - W = meaningful utterance
   - M/V = utterance of mother/father

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.
Japanese-Ota

Project Name: Japanese-Ota

Investigator(s): Mitsuhiko Ota

Contact: mits@ling.ed.ac.uk

Location: Washington, DC

Number of Participants: 3

Nature of study: naturalistic, longitudinal

Media Type (if any): Audio cassette tapes, and video super 8 cassette tapes

Citation Information:

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiromi</td>
<td>1982-12-09</td>
<td>1;00.21 – 2;00.07</td>
<td>22</td>
<td>F</td>
</tr>
<tr>
<td>Kenta</td>
<td>1982-05-15</td>
<td>1;05.18 - 2;06.07</td>
<td>26</td>
<td>M</td>
</tr>
<tr>
<td>Takeru</td>
<td>1982-03-26</td>
<td>1;04.23 – 2;00.19</td>
<td>15</td>
<td>M</td>
</tr>
</tbody>
</table>

The participants were from a Japanese community in the Washington, D.C. area. All children were at the beginning of the one-word period and living in a family that spoke only Japanese among themselves. The children had varying degrees of exposure to English.

The data were collected through a series of recordings, in two to three week intervals, of spontaneous utterances rather than by controlled elicitation techniques. Audio recordings were made on standard cassette tapes with a Sony WM-D6C analog tape recorder and a Sony ECM-909A microphone. There were also video-recordings made on super 8 cassette tapes with a Samsung SCX915 camcorder.

The data were transcribed using a standard Romanization system for Japanese and IPA. The target adult form was determined from the context of the utterance. Indeterminable utterances were removed from the analysis. Spontaneous speech and imitations of the adult's utterances were coded for. In the case of imitations, the full adult utterance is provided.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by the above reference.
Project Name: Japanese-Stanford
Investigator(s): Marilyn Vihman
Contact: mv509@york.ac.uk
Location: Stanford, CA
Number of Participants: 6
Nature of study: naturalistic, longitudinal
Media Type (if any): audio cassettes and/or reel-to-reel/VHS video

Citation Information:

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>TBA</td>
<td></td>
<td>2</td>
<td>TBA</td>
</tr>
<tr>
<td>emi</td>
<td>TBA</td>
<td>0;10.14 – 1;04.07</td>
<td>4</td>
<td>F</td>
</tr>
<tr>
<td>har</td>
<td>TBA</td>
<td>1;01.14 – 1;06.07</td>
<td>5</td>
<td>M</td>
</tr>
<tr>
<td>kaz</td>
<td>TBA</td>
<td>1;00.17 – 1;03.15</td>
<td>5</td>
<td>F</td>
</tr>
<tr>
<td>ken</td>
<td>TBA</td>
<td>1;00.04 – 1;05.26</td>
<td>6</td>
<td>M</td>
</tr>
<tr>
<td>tar</td>
<td>TBA</td>
<td>1;02.07 – 1;11.02</td>
<td>1</td>
<td>M</td>
</tr>
</tbody>
</table>

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.
The characters for the target vowel in the WorldBet and the WorldBetb.seg tiers differ because the representation in the WorldBet tier denotes the vowel phoneme whereas the representation in the WorldBet.seg tier designates the (potentially larger) "vowel class" that we set up to be able to compare the effects of coarticulatory context on initial consonant accuracy across the four languages.

Here are the five groups that we set up for English:

"I" includes both long tense [i:] and short lax [I] (so, e.g., "ti" is elicited in <teepee> and <teacher> and also in <tickle>)

"E" includes both long tense [el] and short lax [E] (e.g., "te" is elicited in <taste> and <tail> and also in <tent>)

"A" includes both (or all three of) low back [A] (and low-mid back rounded [O], which does not contrast with [A] in most Ohio dialects) and low-mid [a] (e.g., "ta" is elicited in <taco> and <tall> and <tongue>)

"O" is just the long tense [oU]

"U" is both [u:] and [U]

Now, for the vowel classes, in Cantonese we had seven, because we wanted to be able to compare the contrastively front rounded vowels of Cantonese to the analogous classes in French (e.g., to contrast the effect of "u" versus "y" on accuracy of the preceding dental consonants, where the high back vowel is phonotactically illegal in Cantonese). Also, in Cantonese, the tones have spaces after them. The following are the seven vowel classes that were set up for Cantonese:

"i" includes long [i:] both as a simple vowel and as the nucleus of a diphthongs, so, e.g., "khi" is elicited in the morpheme "khi:u23" as well as in morphemes "khi:n21" and "khi:t3" (I'm writing just
the first zi in all three words)

"e" includes both long [E:] and short [e] and the latter both as a simple vowel and as the nucleus of the diphthong [ei]

"a" includes both the short low central vowel "ax" (in WorldBet" and the long low back rounded vowel that should be "5: in WorldBet, but which we’re writing as the easier to read "a:" and it includes both of these vowels as both as simple vowels and as the nuclei of diphthongs

"o" includes both the short high-mid [o] (as the simple vowel or as the nucleus of the diphthong [ou]) and the long low-mid [>:] (This “>” is how I should have designated the analogous vowel in English which is included in the English "a" class instead, since in most Ohio dialects it doesn’t contrast with [A], so here’s a place where it was impossible to set up optimal vowel class comparisons.)

"u" includes long [u:] both as a simple vowel and as the nucleus of [u:i]

"oe" includes both the long low-mid rounded front vowel that is called "8" in WorldBet and the mid-central rounded vowel (IPA "squashed theta") as the nucleus of the non-back rounded diphthong "oxy"

"y" includes only the long high front rounded vowel [y:]

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

---

**Portuguese-Lisbon**

**Project Name:** EP_Mono

**Investigator(s):** Susana Correia  
Teresa da Costa

**Contact:** correia.smd@gmail.com  
teresadacosta@gmail.com

**Location:** Lisbon and area, Portugal

**Number of Participants:** 5

**Nature of study:** naturalistic, spontaneous, longitudinal, monolingual

**Media Type (if any):** QuickTime movie files (.mov)

**Citation Information:**

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clara</td>
<td>01-Jun-2005</td>
<td>0;11 - 1;10</td>
<td>12</td>
<td>F</td>
</tr>
<tr>
<td>Inês</td>
<td>19-Nov-1992</td>
<td>0;11 - 4;2</td>
<td>30</td>
<td>F</td>
</tr>
</tbody>
</table>
This corpus contains data from 5 European Portuguese monolingual children during 134 recording sessions. The children’s speech was recorded every other week in the case of João and Luma, and monthly in the case of Inês, Clara and Joana. The data were recorded at the children’s homes, during spontaneous speech. The mother and/or caretaker and the researcher were present during the recording sessions.

The data from Clara, João and Luma were recorded in a Canon MV210 digital video camcorder and videotaped in analogue format (Hi8 cassettes), between 2004 and 2007. Inês and Joana were recorded on a Sony Handycam Video 8, AF Hi-Fi Stereo, during the 1990’s. In all cases, the analogue format was imported to digital format using iMovie© software, and compressed for 320x240 (.mov) format.

The data were manually entered into the Phon application. Orthographic and phonetic transcriptions were made of the target and children's actual forms. The sessions were transcribed and exchanged between two transcribers for revisions. Each transcriber was responsible for transcribing 50% of the children’s speech and revising the other 50%. In dubious cases, a third researcher carried out a blind transcription and, in the case of persistent indecision, that utterance was marked with an asterisk (*) as unintelligible.

All of the specific research questions formulated within this project were related to the development of segmental properties (place and manner features) in Portuguese children. The analytic tools available through the Phon application enabled the researcher to extract information on the general order of acquisition of the EP consonantal inventory and the segmental shape of early words. The analytic features used in the study were: (i) searching and analyzing consonantal place and manner features in onset position; (ii) discriminating between homorganic and non-homorganic feature combinations at the word-level; (iii) searching and analyzing consonantal place and manner features, discriminating between stressed and unstressed syllables; (iv) searching/analyzing consonant harmony and metathesis in children’s speech.

The research questions underlying this project deal with the identification of and order of acquisition for primary word stress and the study of the early shape of prosodic words in Portuguese children. The analytic features used in this study were: (i) searching for and analyzing the number of words uttered by the children; (ii) searching for and analyzing the number of syllables within a word; (iii) searching for and analyzing the syllable structure in different word positions; (iv) searching for and analyzing stress patterns.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.
Portuguese-Freitas

Project Name: Portuguese-Freitas
Investigator(s): Maria João Freitas
Contact: joaofreitas@fl.ul.pt
Location: Lisbon area, PT
Number of Participants: 7
Nature of study: naturalistic, longitudinal & cross-sectional samples
Media Type (if any): Video with Hi-Fi stereo

Citation Information:

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inês</td>
<td>1992-11-19</td>
<td>0;11.13 – 1;10.29</td>
<td>12</td>
<td>F</td>
</tr>
<tr>
<td>João</td>
<td>1992-03-29</td>
<td>0;10.02 – 2;08.27</td>
<td>23</td>
<td>M</td>
</tr>
<tr>
<td>Laura</td>
<td>1991-03-20</td>
<td>0;02.29 – 3;03.10</td>
<td>12</td>
<td>F</td>
</tr>
<tr>
<td>Luís</td>
<td>1991-12-23</td>
<td>1;09.29 – 2;11.02</td>
<td>12</td>
<td>M</td>
</tr>
<tr>
<td>Marta</td>
<td>1992-11-03</td>
<td>1;02.00 – 2;02.17</td>
<td>12</td>
<td>F</td>
</tr>
<tr>
<td>Pedro</td>
<td>1991-03-27</td>
<td>2;07.00 – 3;07.24</td>
<td>12</td>
<td>M</td>
</tr>
<tr>
<td>Raquel</td>
<td>1991-06-01</td>
<td>1;10.02 – 2;10.08</td>
<td>10</td>
<td>F</td>
</tr>
</tbody>
</table>

The recordings for this project were made in the child's home, normally in the child's bedroom. Naturalistic data were collected both longitudinally and cross-sectionally. Each child was videotaped for a period of one year longitudinally (João was videotaped for a period of 2 years). The age of recordings for all sessions were from 0;10 onward and 3;7 onward.

Each recording session lasted from 30 to 60 minutes. The recordings were made using a Sony Handycam video 8, AF Hi-Fi Stereo. Digital recordings will be made available through Centro de Linguística da Universidade de Lisboa by December, 2010.

The data collection was supported by Fundação para Ciência e a Tecnologia, research project PCSV/C/LIN/524/93. The data transcription was first compiled within the CHILDPHON database, developed at the Max-Planck Institut for Psycholinguistics in 1990 and first used by Fikkert (1994) and Levet (1994).

All sessions from João, Inês and Marta were transcribed in full. All sessions from Luís, Raquel, Laura and Pedro were partially transcribed from the 10 minute to the 25 minute mark in each session. Transcriptions were performed by a native speaker of European Portuguese. All problematic transcriptions were noted and listened to by at least two other judges. In case of remaining doubts on transcription, the utterance was not included in the database.

Due to the goals of the original research project, the corpus represents only isolated words or sequences containing external sandhi phenomenon.
See Costa 2010 and Correia 2010 for full access to the Inês corpus, and for documentation of early stress patterns. Transcription of stress in reduplicated forms was not revised in this corpus.

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

---

**Romanian-Kern**

**Project Name:** Romanian-Kern  
**Investigator(s):** Sophie Kern  
**Contact:** Sophie.Kern@univ-lyon2.fr  
**Location:** Bucharest, RO  
**Number of Participants:** 4  
**Nature of study:** naturalistic, longitudinal, monolingual  
**Media Type (if any):** digital video (.mov, .mp4)

**Citation Information:**

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>2000-11-08</td>
<td>0;07.04 – 2;00.04</td>
<td>31</td>
<td>F</td>
</tr>
<tr>
<td>Anna</td>
<td>2000-11-22</td>
<td>0;08.11 – 1;02.10</td>
<td>10</td>
<td>F</td>
</tr>
<tr>
<td>Matei</td>
<td>2000-11-09</td>
<td>0;08.10 – 2;01.11</td>
<td>34</td>
<td>M</td>
</tr>
<tr>
<td>Vlad</td>
<td>2000-09-08</td>
<td>0;08.28 - 2;00.27</td>
<td>28</td>
<td>M</td>
</tr>
</tbody>
</table>

Five types of data were collected. First, one hour of spontaneous vocalization data was audio and video recorded every two weeks from 8 months of age through 25 months of age. Recording took place in children’s homes. The parents were told to follow their normal types of activities with their child. Second, minimally 1,000 dictionary entries from the ambient language employed by the parents of each child participant were analyzed for comparison with the child data for that language. Parental reports were administered using adaptations of the MacArthur Development Inventories (Fenson et al., 1993) respectively elaborated for Dutch for French participants. Mothers filled out the questionnaire once in a month. For the remaining languages there is no
adaptation yet, but one could imagine using the spontaneous data to elaborate the same instrument. An object manipulation categorization task was administered every two months. This task was conceived to evaluate the children's spontaneous nonverbal categorization abilities. Several toys, which were consistent across the language groups served as stimuli. Each task involved a contrast of objects from two different categories (animal, means of locomotion, furniture).

Children were developing normally by community standards as well as reports from parents and physicians regarding developmental milestones were observed in the normal daily environment. These children were becoming monolingual speakers of French, Romanian and Tunisian Arabic. Languages were chosen to include different language families as well as for the contrasts they show in phonetic and phonological features of interest for early language development. These characteristics include word length, syllabic types as well as phonemic inventory diversity.

One hour of spontaneous vocalization data was audio and video recorded every two weeks from 8 months of age through 25 months of age. Recording took place in children's homes. The parents were told to follow their normal types of activities with their child. After collection, data was phonetically transcribed using the International Phonetic Alphabet (IPA). Broad phonetic transcriptions were used, supplemented by some diacritics (mainly for palatalized, pharyngealized, nasalized sounds and duration of sounds). Tokens considered as single utterance strings were bounded by one second of silence, noise or adult speech.

This work was supported by the EUROCORE Program “The Origin of Man, Language and Languages” (OMLL) and the French CNRS program “Origine de l’Homme, du Langage et des Langues” (OHLL).

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

---

**Spanish-PAIDUS**

**Project Name:** Spanish-PAIDUS  
**Investigator(s):** Antonio Maldonado, Conxita Lleó  
**Contact:** Universidad Autónoma de Madrid, lleo@uni-hamburg.de  
**Location:** Madrid, ES  
**Number of Participants:** 5  
**Nature of study:** naturalistic, longitudinal/cross-sectional, babbling, first words, informal testing  
**Media Type (if any):** Wave files  
**Citation Information:**
Lleó, Conxita, Michael Prinz, Christliebe El Mogharbel & Antonio Maldonado (1996). Early phonological acquisition of German and Spanish: A reinterpretation of the continuity


<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ines</td>
<td>1991-02-23</td>
<td>0;09.27 – 1;06.12</td>
<td>7</td>
<td>F</td>
</tr>
<tr>
<td>Jose</td>
<td>1991-06-17</td>
<td>0;09.16 – 2;06.12</td>
<td>21</td>
<td>M</td>
</tr>
<tr>
<td>Juan</td>
<td>1990-12-30</td>
<td>0;09.17 – 2;01.13</td>
<td>11-12 (Bruder)</td>
<td>M</td>
</tr>
<tr>
<td>Maria</td>
<td>1990-10-17</td>
<td>0;09.15 – 2;07.17</td>
<td>16</td>
<td>F</td>
</tr>
<tr>
<td>Miguel</td>
<td>1990-08-13</td>
<td>0;10.04 – 3;01.28</td>
<td>22</td>
<td>M</td>
</tr>
</tbody>
</table>

The project BIDS/PAIDUS was carried out at the University of Hamburg (Department of Romance Languages) with the support of the Deutsche Forschungsgemeinschaft under the direction of Conxita Lleó. The project began in 1990 first as BIDS, with the goal of comparing the phonetic characteristics of babbling in German and comparing it with babbling in Spanish. With this purpose, a collaboration with Antonio Maldonado of the Universidad Autónoma de Madrid (Dept. of Psychology) was started and we applied for a joint grant from “Acciones Integradas”, so that...
the two teams could visit each other. The German data were collected in Hamburg, Germany, and the Spanish data in Madrid, Spain.

All five German children had been born in Hamburg and lived there. Their parents were all German and spoke the Northern variety of German. The children were visited from very early on, before they began to produce their first words, as we wanted to study their babbling. Some informal tests, in order to figure their (passive) vocabularies were conducted. After the children began saying words, we continued recording them within the project PAIDUS, which had as its goal to finding out how phonological parameters were possibly fixed for German as opposed to Spanish. The German children were visited every 2 weeks at first, and later on once a month. Recordings were made at the children's homes by two research assistants of the project. At first, the mother used to be present, but later on, we made most of the recordings with the child alone interacting with one research assistant. We tried to record the child under spontaneous conditions, playing, chatting, etc. We also brought some objects (little animals or toys) in order to elicit certain words from all children, making their phonetic and phonological characteristics comparable. The mother tried to keep a record of possible new words that were new in the child's vocabularies, and we tried to elicit them, as well. We also used books with pictures, which belonged to the child and others that we brought. The recordings were audio recordings, and in order not to miss the ambient setting and context, one of the researchers was writing down all possible details, including what the child was doing, while the other assistant was interacting with the child.

Both projects, BIDS and PAIDUS, were financed by the DFG with grant to Conxita Lleó. BIDS from 1989 until 1991 and PAIDUS from 1991 until 1994. Research Assistants in the project BIDS were Christliebe El Mogharbel and Kerstin Feuge. Research Assistants in the project PAIDUS were Christliebe El Mogharbel and Michael Prinz.

Researchers using the BIDS/PAIDUS data should send a copy of their publication based on the data to the contributor of the data. In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

---

**Swedish-Lacerda**

Cornelia was born 23-SEP-2004. Ebba was born 02-SEP-2004. Teresa was born 17-AUG-2004.

<table>
<thead>
<tr>
<th>Session</th>
<th>Date</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>c06</td>
<td>2005-OCT-10</td>
<td>1;0.17</td>
</tr>
<tr>
<td>c08</td>
<td>2005-DEC-05</td>
<td>1;3.02</td>
</tr>
<tr>
<td>c09</td>
<td>2006-JAN-13</td>
<td>1;3.21</td>
</tr>
<tr>
<td>c10</td>
<td>2006-FEB-17</td>
<td>1;4.25</td>
</tr>
<tr>
<td>c12</td>
<td>2006-DEC-18</td>
<td>2;2.25</td>
</tr>
<tr>
<td>c13</td>
<td>2007-MAY-07</td>
<td>2;7.14</td>
</tr>
<tr>
<td>e05</td>
<td>2005-OCT-10</td>
<td>1;1.08</td>
</tr>
<tr>
<td>e06</td>
<td>2005-NOV-08</td>
<td>1;2.06</td>
</tr>
<tr>
<td>e07</td>
<td>2006-JAN-16</td>
<td>1;4.14</td>
</tr>
<tr>
<td>e08</td>
<td>2006-FEB-06</td>
<td>1;5.04</td>
</tr>
</tbody>
</table>
Sound files are named for session. There were separate WaveSurfer transcript files for the child, caregiver, and the investigator and these were merged to create the CHILDES transcripts.

---

**Swedish-Stanford**

**Incomplete**

*Project Name:* Swedish-Stanford  
*Investigator(s):* Marilyn Vihman  
*Contact:* mv509@york.ac.uk  
*Location:* Stockholm, SE  
*Number of Participants:* 4 (or 5)  
*Nature of study:* naturalistic, longitudinal  
*Media Type (if any):* audio cassettes and/or reel-to-reel/VHS video

**Citation Information:**  

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>cur</td>
<td>TBA</td>
<td>0;09.03 – 1;04.05</td>
<td>TBA</td>
<td>M</td>
</tr>
</tbody>
</table>
In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.

<table>
<thead>
<tr>
<th>Participant Name</th>
<th>Date of Birth</th>
<th>Age range</th>
<th>Number of Sessions</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fereyel</td>
<td>2002-04-09</td>
<td>0;08.09 – 2;00.21</td>
<td>26</td>
<td>F</td>
</tr>
<tr>
<td>Iyed</td>
<td>2002-01-30</td>
<td>0;08.22 - 2;00.00</td>
<td>30</td>
<td>F</td>
</tr>
<tr>
<td>Malek</td>
<td>2002-02-24</td>
<td>0;10.11 – 2;00.04</td>
<td>24</td>
<td>M</td>
</tr>
<tr>
<td>Zaidaan</td>
<td>2002-01-24</td>
<td>0;04.15 – 2;00.10</td>
<td>31</td>
<td>F</td>
</tr>
</tbody>
</table>

Five types of data were collected. First, one hour of *spontaneous vocalization* data was audio and video recorded every two weeks from 8 months of age through 25 months of age. Recording took place in children's homes. The parents were told to follow their normal types of activities with their
child. Second, minimally 1,000 *dictionary* entries from the ambient language employed by the parents of each child participant were analyzed for comparison with the child data for that language. *Parental reports* were administered using adaptations of the MacArthur Development Inventories (Fenson et al., 1993) respectively elaborated for Dutch for French participants. Mothers filled out the questionnaire once a month. For the remaining languages there is no adaptation yet, but one could imagine using the spontaneous data to elaborate the same instrument. An object manipulation *categorization task* was administered every two months. This task was conceived to evaluate the children’s spontaneous nonverbal categorization abilities. Several toys, which were consistent across the language groups served as stimuli. Each task involved a contrast of objects from two different categories (animal, means of locomotion, furniture).

Children were developing normally by community standards as well as reports from parents and physicians regarding developmental milestones were observed in the normal daily environment. These children were becoming monolingual speakers of French, Romanian and Tunisian Arabic. Languages were chosen to include different language families as well as for the contrasts they show in phonetic and phonological features of interest for early language development. These characteristics include word length, syllabic types as well as phonemic inventory diversity.

One hour of spontaneous vocalization data was audio and video recorded every two weeks from 8 months of age through 25 months of age. Recording took place in children’s homes. The parents were told to follow their normal types of activities with their child. After collection, data was phonetically transcribed using the International Phonetic Alphabet (IPA). Broad phonetic transcriptions were used, supplemented by some diacritics (mainly for palatalized, pharyngealized, nasalized sounds and duration of sounds). Tokens considered as single utterance strings were bounded by one second of silence, noise or adult speech.

This work was supported by the EUROCORE Program “The Origin of Man, Language and Languages” (OMLL) and the French CNRS program “Origine de l’Homme, du Langage et des Langues” (OHLL).

In accordance with CHILDES rules, any use of data from this corpus must be accompanied by at least one of the above references.