This is a guide to CHILDES data on the acquisition of Slavic languages. For a general introduction to the CHILDES database, please consult intro.pdf. The links in the table below are clickable, as are the thumbnails to the left.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Age Range</th>
<th>N</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Croatian – Kovacevic on page 2</td>
<td>1;3–2;8; 1;10–2;11; 0;10-3;2</td>
<td>3</td>
<td>Two girls and a boy learning Croatian in Zagreb.</td>
</tr>
<tr>
<td>Polish – CDS on page 5</td>
<td>1-8</td>
<td>Many</td>
<td>Frequency list for child-directed speech from eight corpora</td>
</tr>
<tr>
<td>Polish – Szuman on page 2</td>
<td>1;5–7;9</td>
<td>10</td>
<td>Diary data collected by Szuman and his students and computerized by Magdalena Smoczynska</td>
</tr>
<tr>
<td>Polish – Weist-Jarosz on page 9</td>
<td>1;7–2;6</td>
<td>3</td>
<td>Three children learning Polish with six files for each child</td>
</tr>
<tr>
<td>Russian – Protassova on page 12</td>
<td>N/A</td>
<td>1</td>
<td>Longitudinal study of a child learning Russian</td>
</tr>
<tr>
<td>Russian – Tanja on page 13</td>
<td>2;5–2;11</td>
<td>1</td>
<td>A child learning Russian in a monolingual environment in the United States</td>
</tr>
<tr>
<td>Serbian on page 15</td>
<td>1;6-4;0</td>
<td>8</td>
<td>Recordings in homes with many people included</td>
</tr>
<tr>
<td>Slovenian – Kagar on page 16</td>
<td>5;0</td>
<td>20</td>
<td>Arguments patterns in kindergarten children</td>
</tr>
</tbody>
</table>
1. Croatian – Kovacevic

Melita Kovacevic
University of Zagreb
Department of Language & Speech Pathology
Laboratory for Psycholinguistic Research
V. Holjevca 20, 10000 Zagreb, Croatia
labpolin@erf.hr
melita.kovacevic@public.srce.hr

Acknowledgements:
When using Antonija and Marina corpora, please cite the title of the project and the name of head of the project: Acquisition of Croatian in Crosslinguistic Perspective, Kovacevic, M. (2003).

Restrictions:
By contributing our data to the CHILDES system we do not impose any particular restrictions on the use of the data. However, we would appreciate if researchers send the copies of articles that make use of the data or send us a reference.

Warnings:
As Croatian keyboard has been used, there might be some problems with Croatian characters, e.g. š, d, ž, č and ě. The files were checked with the CHECK program and have passed it successfully. However, some inconsistencies in encoding were inevitable due to the fact that a few different transcribers worked on the transcripts of each child over a long period of time. For all questions please write to: Marijan Palmovic (palmovic@erf.hr) or Gordana Hrzica (ghrzica@erf.hr).

Pseudonyms:
The informed consents for the use of children data have been obtained. There are no pseudonyms, but only the first names are used.

History:
The data collection started in 1993 when the first project on the acquisition of Croatian was initiated. From that point up to now the project has been renewed and the new grant has been rewarded in 2002 for the three year period. The project is headed by Prof. Melita Kovacevic, Department of Speech & Language Pathology, University of Zagreb (melita.kovacevic@public.srce.hr) and it has been financed by the Croatian Ministry of Science and Technology (Project No. 013002).

The Project follows language development of the total of ten children. Among them there are two pairs of twins, where one pair is late talkers. While some corpora are already completely coded, some are still in the process of recording. Parents of all children included in our study expressed interest and wish their children language development to be followed. Most of the parents are in some way connected to language fields, e. g. language pathology, language teaching, and linguistics.
Antonija's data are transcripts of the audio-recordings collected during 1994 - 1996 in Zagreb. Antonija was recorded in her home in spontaneous interactions with her parents and grandparents. The period of about one month, between 1;7 and 1;9, is missing due to family reasons. Each recording session lasted about 45 minutes and usually there were three sessions a month (see the table). The recordings and the transcriptions were made by Antonija's mother, Drazenka Blazi. Additional transcription and encoding was done by Maja Andjel, Antigona Katicic and Marijan Palmovic. All the transcribers are linguists and trained for the transcriptions. M. Palmovic and M. Andjel are research assistants (University of Zagreb) and A. Katicic is a Ph.D. student (University of Vienna).

Marina's data are transcripts from audio-recordings collected during 1994 and 1995. They cover the period from Marina's age of 1;10 to 2;11 with three months missing (2;4, 2;10 and 2;11) due to the escalation of war in Croatia at the time. The child has been recorded for a longer period of time, but all the files are not in the stage to be released. Since some of the recordings have been done outside, in a yard, the quality of sound is not satisfactory and as such requires additional work. The present files were partially transcribed by Blazenka Brozovic, research fellow, Petra Makanec, student, and Maja Andjel. Final checking of the transcriptions and addition of the error codes has been done by Marijan Palmovic in the Laboratory for Psycholinguistic Research, Dept. for Speech & Language Pathology, University of Zagreb (labpolin@erf.hr).

Vjeran's data are transcripts of audio-recordings collected from 1995 to 1997. All the recordings were done in spontaneous interactions, mostly in his home in Zagreb, either with his parents or with person who was recording. Some of the recordings were primarily based on picture book readings, and this should be taken into account when performing analysis. Some recordings took place outside home, at the playground or the neighboring store. The recordings covered the period from Vjeran's age of 0;10 to 3;02. Each recording session lasted about 45 minutes, 3 times per month on average (see the table). The recordings were mainly done by Blazenka Brozovic, research fellow, and some have been done by his mother. Transcriptions were done by Blazenka Brozovic and Jelena Kuvac, research assistant. Final checking of the transcripts and some additional coding has been done by Marijan Palmovic and Gordana Hrzica in the Laboratory for Psycholinguistic Research, Dept. for Speech & Language Pathology, University of Zagreb (labpolin@erf.hr). All the transcribers have been trained to use CHILDES transcription and coding system.

Codes:
For the purposes of the transcription a new code was added to the depfile - @x:syl. During the sessions it was quite common that Marina's mother elicited a word by uttering the first syllable of the word (and other syllables if necessary). Marina often responded in just adding the next syllable or the rest of the word. Regarding Antonija's files, rather than producing a single letter, Antonija was often producing a whole syllable. The new code was needed since these Antonijas' utterances neither were babbling nor incomplete words.
Biographical data:

Antonija
Antonija's parents are middle class urban dwellers: her mother is an university teacher (speech & language pathologist) and her father is an engineer. Both parents are born and raised in Zagreb. They speak Zagreb Stokavian dialect. Croatian is marked with a number of dialects. The differences among these dialect groups could be more substantial than the differences that divide standard languages within the Slavic language family. In Zagreb (Croatian capital) there are two major dialects, so called Zagreb Stokavian - to be closer to the standard Croatian - and Zagreb Kajkavian - to be closer to Kajkavian dialect. The dialects differ on all language components levels: lexicon, phonology, morphology and syntax. Being an university teacher and a speech pathologist, Antonija's mother pays much attention to the way she speaks to her child often repeating what the child has just said. However, in Antonija's interactions with her grandparents strong elements of Zagreb Kajkavian dialect can be observed - the most noticeable being the question word kaj (eng. *what*). It should be noted that Antonija lives in a three generation household in Zagreb suburb, being all the time with both of the grandparents as well. Kajkavian influence can also be heard in prosody as a lack of the two rising accents, but this cannot be seen in the transcripts since such information was neglected not being a focus of the study. At the time of the recordings Antonija was the only child.

Marina
Marina's parents are middle class urban dwellers as well: her mother is an assistant professor at the Department for Speech & Language Pathology, University of Zagreb and her father is an actor in a Croatian National Theatre in Zagreb. Her mother was born and raised in Zagreb, while her father came from Dalmatian coast (Zadar) to study in Zagreb. They both speak Zagreb Stokavian dialect with some elements of Zagreb Kajkavian although her father comes from Cakavian dialect region (another distinct dialect of Croatian). However, being a trained actor, his everyday speech is closed to standard. At the time of the recordings Marina had a sibling - infant girl, Vita. She has been only mentioned during the recording sessions.

Vjeran
Vjeran’s parents also belong to middle class: his mother is English teacher working in a private college and his father is an electro engineer. His mother also received a degree in phonetics, which made her quite sensitive for different language/speech issues. Both parents were born and raised in Zagreb. They speak Zagreb Stokavian dialect (please see the above information on the dialect). Vjeran’s babysitter, with whom he was spending a few hours daily, was a woman who came to Zagreb from another Croatian region, Lika. The speech of that region is Stokavian dialect which has been preserved in the speech of the babysitter. This dialect is the base of Croatian standard. Vjeran does not have any siblings and was mostly surrounded by adults at the time when the recordings had been done.
2. Polish – CDS

Ewa Haman¹, Bartłomiej Etenkowski², Magdalena Łuniewska¹, Joanna Szwabe³, Ewa Dąbrowska⁴, Marta Szreder⁵, Marek Łaziński¹
¹University of Warsaw, Warsaw, Poland; ²Charles University of Prague, Prague, Czech Republic; ³Adam Mickiewicz University, Poznań, Poland; ⁴Northumbria University, Newcastle, UK; ⁵University of York, York, UK

The Polish Frequency List of Child-Directed Speech is based on utterances of parents, grandparents, experimenters and other people (older than 8 years) talking in the presence of the children. Although not all of these utterances are child-directed, they still represent the type of speech that children are exposed to (stories and poems telling, talks between parents). The Polish Frequency List of CDS compiles data from seven corpora of child-directed speech and child language, with the oldest coming from 1940s and the newest coming from 2000s.

The data include: (1) the Polish part of the CHILDES database: Szuman Corpus and Weist Corpus; (2) Polish Child Speech Corpus (Joanna Szwabe, Adam Mickiewicz University, Poland); (3) Gdańsk corpus – speech diary of one girl (Ewa Dąbrowska, Northumbia University, UK & Elena Lieven, Manchester University); (4) speech diary of two brothers (Marta Szreder; York University); (5) narrative mother-child data (Ewa Haman, University of Warsaw & Andrea Zevenbergen, State University of New York, Fredonia); (6) speech diary of twin sisters (Ewa Haman, University of Warsaw).

All data are from naturalistic conversations, except for the narrative mother-child data from the crosslinguistic project of Zevenbergen and Haman (e.g. Zevenbergen, Haman, Olszańska & Thielges, 2008). Those data were included as well, since their task was to talk about any three past events freely chosen from their own experience.

Although only the Szuman and Weist corpora are currently available from CHILDES, all other data are intended to be attached there in the future.

Preparation of the Polish CDS Corpus was supported by the Ministry of Science and Higher Education in Poland (grant no N N106 051436, awarded for 2009-2012 to Ewa Haman).

The detailed information about the data is shown in the Table 1.
Table 1

<table>
<thead>
<tr>
<th>Author</th>
<th>Dates of registries</th>
<th>No of children</th>
<th>Children’s age</th>
<th>Children’s gender</th>
<th>All word tokens</th>
<th>Word tokens in CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Szuman</td>
<td>1945-46, 1953-61, 1964-65</td>
<td>12</td>
<td>0;10 - 6;11</td>
<td>female, male</td>
<td>1 035 691</td>
<td>375 168</td>
</tr>
<tr>
<td>Weist</td>
<td>1980-1981</td>
<td>4</td>
<td>1;7 - 3;2</td>
<td>female, male</td>
<td>73 745</td>
<td>44 003</td>
</tr>
<tr>
<td>Dąbrowska</td>
<td>1999</td>
<td>1</td>
<td>2;0 - 2;1</td>
<td>female</td>
<td>101 529</td>
<td>68 833</td>
</tr>
<tr>
<td>Haman</td>
<td>2002-2003</td>
<td>2</td>
<td>1;11 - 2;23</td>
<td>female</td>
<td>105 280</td>
<td>72 740</td>
</tr>
<tr>
<td>Haman &amp; Zevenbergen</td>
<td>2006-2007</td>
<td>45</td>
<td>3;1 - 6;1</td>
<td>female, male</td>
<td>50 393</td>
<td>36 265</td>
</tr>
<tr>
<td>Szreder</td>
<td>2008-2009</td>
<td>2</td>
<td>2;2 - 3;7</td>
<td>male</td>
<td>29 424</td>
<td>16 641</td>
</tr>
<tr>
<td>Szwabe</td>
<td>2006-2009</td>
<td>62</td>
<td>3;0-6;11</td>
<td>female, male</td>
<td>383 429</td>
<td>185 789</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>128</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>1 779 491</strong></td>
<td><strong>799 439</strong></td>
</tr>
</tbody>
</table>

As the Table 1 shows, our data include transcriptions of speech directed to 128 children aged 0;10-6;11. The data include more than 799 thousands word tokens in CDS.

The Polish language uses a high degree of inflection. It has seven cases and three genders. Declination endings in Polish depend on case (nominative, genitive, dative, accusative, locative, instrumental, vocative), number (singular or plural), gender (masculine, feminine, neuter), animacy (animate or inanimate) and whether a particular word denotes a human or not. Thus one Polish word may have more than ten forms. For example, adjective *góły* (bare/naked) can be used in a sentence in 11 ways: *góły, gołego, gołym, gola, golą, golej, gole, goli, gołych, gołymi* marking 70 inflectional forms – 7 cases x 2 numbers x 5 genders (comparative and superlative forms are not listed). Because we wanted to get e.g. information about the variety of words in CDS (not of their forms) it was necessary to prepare a frequency list that counts all the forms of the same word together and contains the basic form of the word as well as inflected forms. (E.g. *góły* 46 as well as: *gola 16, goli 14, gole 9, goly 6, gołego 1*). The Polish Frequency List of CDS is prepared in this way.

A database (PostgreSQL 8.3) comprising all data available in CHAT format was prepared to get the frequency list of CDS in the data. The database contains all corpora except for Polish Child Speech Corpus which is in TEI format. The database reflects the structure of CHAT format. Additionally each utterance has codes assigned for its author, listeners, their age and gender. The database structure makes it possible to get the utterances and the words from the whole corpus that fulfill the specified criteria as the author’s age and/or gender, the listeners’ age and/or gender, the registering date of the utterance. It is also possible to merge more criteria together. It is e.g. possible to get the frequency list of speech of people older than 10 years directed to two-year-old girls or the frequency list of speech of adult men directed to boys aged between 5 and 7. A first raw frequency list for all CDS data available in CHAT format was obtained from the

---

1 Although Szuman Corpus includes also transcriptions of speech directed to child aged 7;0-10;7 they come from talks with only one child. Because other include no transcriptions of speech directed to children older than 7;0 we decided to ignore this part of Szuman Corpus.
database. The raw frequency list of CDS in Polish Child Speech Corpus was prepared by its author Joanna Szwabe. The corpora (except of the Polish Child Speech Corpus) were inserted into the database. We got the utterances that fulfill the specified criteria (author older than 8;0, at least one of listeners younger than 7;0) and lower-cased them. From these utterances we got the first raw frequency list. We merged it with Polish Child Speech Corpus raw frequency list which was generated outside the database (by its author Joanna Szwabe) and was already lower-cased. The merged raw frequency list contained all word forms not classified according to the Polish inflection system. All these data (two merged frequency lists) were lemmatized by lemmatizing program prepared by Jarosław Strojek on the base of the Polish Language Corpus (PWN, 2009). Additionally, because of the relatively high number of homonyms in Polish, frequencies of homonymic forms were divided according to the homonym list which contains information on the proportion of the distribution of each homonymic form in the Polish Language Corpus (PWN, 2009). Only the homonymy of different lemmas has been accounted, so is the string goli occurring altogether 19 times classified 14 as a form of the lemma goły and 5 times of the lemma golić (się) (shave). The homonymy of two word forms inside one lemma, as gołym: the instrumental singular or the dative plural of goly does not upset our statistics.

All corpora used include more than 1,179,000 word tokens with more than 794,000 word tokens in CDS (speech directed to children aged between 0;10 and 6;11), about 44,000 word types, and 21,000 different lexemes. The 46 most frequent lexemes cover more than half of the CDS items in the corpora and 90% of the CDS items are covered by the first 1,811 lexemes.

Lemmatized list of Polish CDS is available in xls format. Columns show:

A: Inflected forms
B: Inflected forms occurrences (frequencies)
C: Base forms of word
D: Base forms occurrences (frequencies)

There are some fractions in the column showing occurrences. They are the result of rounding the occurrences of the homonyms according the proportion of the distribution of each homonymic form in the Polish Language Corpus (PWN, 2009). For example, for the homonymic form byli the probability of being an inflected form of the word być (to be) is 0.966 and the probability of being an inflected form of the word były (previous) is 0.034. The occurrence of the word byli in our corpora is 68. According to the probabilities above the occurrence of byli as a form of the word być is 95.69 and as a form of the word były is 2.31.

Inquiries should be directed to:
Ewa Haman ewa.haman@psych.uw.edu.pl
Bartłomiej Etenkowski b.etenkowski@gmail.com
Magdalena Łuniewska Madzia.luniewska@gmail.com
3. Polish – Szuman

Magdalena Smoczynska
Department of General Linguistics
Jagiellonian University
Aleja Mickiewicza 9/11
Krakow 31-120 Poland
ulsmoczy@vela.filg.uj.edu.pl

These data were collected by Szuman and his students at the University of Krakow in the years following World War II. In most cases, the children were studied by their parents, many of whom were graduate students in developmental psychology, working with Szuman. In many cases, the public library of Krakow preserves additional materials on this project, including photographs, related treatises by Szuman, and examples of the children’s artwork. Magdalena Smoczynska computerized the Szuman notebooks and contributed her transcripts to CHILDES. The ages of the 10 children are given in the following table.

<table>
<thead>
<tr>
<th>Child</th>
<th>Birthdate</th>
<th>First Session</th>
<th>Last Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basia</td>
<td>12-DEC-1952</td>
<td>1;5</td>
<td>7;9</td>
</tr>
<tr>
<td>Inka</td>
<td>-</td>
<td>0;10.0</td>
<td>7;7</td>
</tr>
<tr>
<td>Jadzia</td>
<td>-</td>
<td>4;1.0</td>
<td>5;1.0</td>
</tr>
<tr>
<td>Janeczek</td>
<td>-</td>
<td>4;2.0</td>
<td>7;3.0</td>
</tr>
<tr>
<td>Jas’</td>
<td>3-JUL-1947</td>
<td>0;11.0</td>
<td>6;6.0</td>
</tr>
<tr>
<td>Kasia</td>
<td>17-DEC-1957</td>
<td>1;3.0</td>
<td>3;11.0</td>
</tr>
<tr>
<td>Krzysz’</td>
<td>14-MAR-1943</td>
<td>2;0.0</td>
<td>3;0.0</td>
</tr>
<tr>
<td>Michal</td>
<td>19-JAN-1949</td>
<td>2;0.0</td>
<td>6;8.0</td>
</tr>
<tr>
<td>Piotrus’</td>
<td>22-FEB-1943</td>
<td>2;0.13</td>
<td>2;9.0</td>
</tr>
<tr>
<td>Tenia</td>
<td>16-APR-1955</td>
<td>1;5.0</td>
<td>2;0</td>
</tr>
</tbody>
</table>

Publications using these data should cite:

4. Polish – Weist-Jarosz

Richard Weist  
Department of Psychology  
SUNY Fredonia  
Fredonia, NY 14063 USA  
weist@al2t.cc.fredonia.edu

Gaja Jarosz  
Department of Linguistics  
Yale University  
New Haven CT USA  
gaja.jarosz@yale.edu

Richard Weist has contributed these data from four children learning Polish. There are six files for each child. The children were tape recorded at the following ages: Marta at 1;7, 1;8, 1;8, 1;9, 1;9, and 1;10.2; Bartosz at 1;7, 1;7, 1;8, 1;8, 1;11, and 1;11; Kubus at 2;1, 2;2, 2;2, 2;4, 2;4, and 2;6; and Wawrzon in 20 sessions from 2;2 to 3;2. All of the children were from middle-class families raised in the urban environment of Poznan, Poland. In general, their parents were highly educated. The children were recorded in their homes (typically an apartment) by two experimenters. One of the experimenters carried a small bag containing the tape recorder and the other took context notes, which were integrated during transcription. In addition to the three child language data sets, we have included a description of the coding. The basic unit of data was a text line, a gloss, and a translation. Context notes are included where available. Because of the use of morphemic glosses, the data are coded morphemically in a way that is very useful for comparative analysis.

This project was supported by NSF BNS 861777, NSF SBR 9309376, NICHD, and the Kosciuszko Foundation. Hanna Wysocka and Katarzyna Witkowska-Stadnik collected Wawrzon's data. Zbigniew Nadstoga and Emilia Konieczna-Tou entered the data. Publications that use these data should cite Weist, Wysocka, Witkowska-Stadnik, Buczowska, and Konieczna (1984) or Weist and Witkowska-Stadnik (1986). We would like to thank Magdalena Smoczynska and Oscar Swan for their helpful comments on our coding system.

Gaja Jarosz supervised the broad phonetic transcription of the data with the help of the open-source Phon software (Rose et al. 2006). The orthographic transcripts were used as the basis for creating phonetic transcriptions of the children’s target pronunciations, and the audio recordings were used to phonetically transcribe the children’s actual productions and align them with the target transcriptions word by word. The transcription of all child productions was first performed independently by two transcribers trained in phonetic transcription, at least one of whom was a native speaker of Polish. Then, two Polish speakers trained in phonetic transcription worked together to create a consensus transcription of all productions, relying on a third phonetically trained native speaker of Polish to adjudicate in cases when agreement could not be reached. The resulting corpus includes phonetic transcriptions of the children’s productions in all the available
audio files, providing word-by-word alignment of target pronunciations and actual pronunciations.

**Boundaries**

We use word groups to delineate phonological word boundaries. We sporadically use ‘#’ to mark various pauses. ‘#’ in the middle of or at the end of an IPA Actual word group indicates a pause inserted by the child. We also use ‘#’ by itself in a word group to indicate an utterance boundary.

In all cases except one, orthographic word boundaries correspond to phonological word boundaries. The only exception is the proclitics ‘w’ [v]/[f] and ‘z’ [z]/[s] which attach to the following word and cannot be pronounced independently. In this case, the orthography tier encodes the orthographic word boundaries, putting the proclitic in its own word group, while the IPA Target and IPA Actual tiers encode the proclitic together with the next word. So for example, ‘[z][kotem]’ would be ‘[[]][skotem]’ on the Target tier and potentially something like ‘[[]][sotem]’ on the Actual tier.

**Tier Conventions**

We have maintained some of the conventions from the original CHAT transcripts of the files and coded them on the orthography tier. In particular, the CHAT transcripts use ‘@x’ codes or ‘*’ to mark special forms, like child-specific forms or lexical errors. When these occurred in the original CHAT transcripts we maintained them in the orthography tier but made sure that the codes occur at the end of word groups. So if the CHAT transcripts applied the code to a sequence of words, we would put the code at the end of each of the words on the orthography tier. For example, something like ‘<mine doggie>*’ would be translated to [mine*][doggie*].

Here are the codes we used:  
- @c a child-specific form  
- @o - onomatopoeia  
- @f - family specific form  
- @s - second language form, or borrowing, or proper name in English (usually)  
- @l indicates lexical error  
- @q (for 'quote') for things the child is reciting from memory or by repetition  
- @i - interjection  
- @w - whispered  
- * or ** after a word indicates a morphological or syntactic error  
- [++] after the final word in an utterance means the utterance was interrupted - put [++] as a separate group at the end  
- [++] utterance was interrupted in the middle of a word - put [++] at the end as a group by itself, and use parentheses to indicate the portion of the unpronounced word  
- [++] at the beginning of an utterance means this is a completion of the adult’s prompt - write [++] at the beginning, and  
- use the code ’@q’ (q for quote) on the words in the middle of a word, - write [++] at the beginning, put parentheses around unpronounced portion of the first word  

CHAT codes that abbreviate special pronunciation modes, such as ‘x5’- indicating five repetitions, were recoded on the orthography tiers (with five separate repetitions).

We put unpronounced (untargeted) portions of the orthographic transcription in parentheses. This sometimes applies to entire words and other times to portions of words that we judged not to be targeted. Portions in parentheses on the orthography tier do not appear on the target tier. If portions of words are omitted for phonological reasons, they remain on the orthography and on the target tiers.

We use [xxx] on the orthography tier as a word group to indicate untranscribable speech (in this case both IPA Actual and IPA Target are left blank).
We use [yyy] on the orthography tier as a word group to indicate speech that we can transcribe but for which we are unable to determine a target pronunciation or target orthography. In this case the IPA Target is blank and the IPA Actual contains the transcription.

IPA Conventions

For the most part each word group is just a sequence of individual IPA symbols that can be treated literally.

One exception is that we've used ligatures for affricates for convenience and to make sure the affricates were consistently differentiated from stop-fricative sequences (which are contrastive in Polish)

We used the postalveolar affricate ligatures for convenience, but these are actually usually transcribed as retroflex and that's what we mean to represent with the postalveolars. They should have the same place of articulation as the (retroflex) fricative series.

Our level of transcription is relatively broad and pretty standard for Polish, but it does encode some non-contrastive phonetic characteristics of the targets and actuals, for example:

- nasal vowels are transcribed on the target tiers according to their standard pronunciation by context (as vowel+nasal stop homorganic with the following noncontinuant consonant and as vowel+nasalized glide otherwise)

- we paid special attention to voicing of obstruents, and target transcriptions account for word-final devoicing and voicing assimilation (including across word boundaries)

due to the longer phonetic length of the palatal portion of palatalized labials (e.g. 'piesek') we coded these palatals as labial-palatal sequences (e.g. [pjɛsɛk]) and we coded the palatalized velars (e.g. 'kiedy') using secondary palatalization (e.g. [kɛdɨ]) in the standard pronunciation and when children produced them as adult-like. These are not contrastive distinctions (there's no contrast between [Cj] and [Cj] in Polish).

Publications using these data should cite:


5. Russian – Protassova

Ekaterina Protassova
Pelimannintie 21-23 F 27
Helsinki 00420 Finland

Ekaterina Protassova of the Russian Academy of Education has contributed data from recordings of her daughter Varvara, born on October 1, 1982, in Moscow, the first and the only child in the family. Her father Alexander (Sasha) was a book illustrator and her mother Ekaterina (Katja) was a psycholinguist. The child was brought up at home. Some days of the week grandparents took care of her, sometimes she spent several hours in a family with two children and a dog. Her grandparents lived at the time in the same flat; both were scientists. The girl’s name is Varvara, which is a Russian equivalent for Barbara. A more common short variant is Varja; diminutives include Varen’ka, Varjusha, Varjunja, and Varjushen’ka. The appellative is Var’, Varjun’, and Varjush. At seven months, Varvara used her first word which was to call herself Ain’ka, so sometimes this name is used by parents.

All of the recordings were taken during 90-minute periods in the usual situations at home or in the summer house by a simple recorder and written down immediately afterwards in Russian. The roman transliteration, English translation, and comments were added in 1995. Childish sound modifications and shortenings of the conventional words are usually included, at least until the fourth session. The dates of the files are as follows:

<table>
<thead>
<tr>
<th>File</th>
<th>Age</th>
<th>File</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1;6.5</td>
<td>4</td>
<td>1;10.14</td>
</tr>
<tr>
<td>2</td>
<td>1;7.13</td>
<td>5</td>
<td>2;0.1</td>
</tr>
<tr>
<td>3a</td>
<td>1;8.24</td>
<td>6</td>
<td>2;4.14</td>
</tr>
<tr>
<td>3b</td>
<td>1;8.24</td>
<td>7</td>
<td>2;10.14</td>
</tr>
</tbody>
</table>
6. Russian – Tanja

Eva Bar-Shalom and
William Snyder
Dept. of Linguistics, U-1145
University of Connecticut
341 Mansfield Road
Storrs, CT 06269-1145 USA

barshalo@uconnvm.uconn.edu
wsnyder@sp.uconn.edu

The Tanja corpus was videotaped and transcribed by Eva Bar-Shalom, a native
speaker of Russian, in collaboration with William Snyder. The project was conducted in
the Child Language Laboratory, Department of Linguistics, University of Connecticut,
and was funded in part by the University of Connecticut Research Foundation. The
corpus contains fifteen longitudinal, spontaneous-speech samples from a monolingual,
Russian-learning girl with the pseudonym “Tanja” between the ages of 2;05.14 and
2;11.20. Tanja was recorded in her home in the United States at a rate of approximately
twice per month. At the time of the study Tanja was an only child, and was cared for at
home by her monolingual (native Russian) mother and her bilingual (native Russian,
ESL) father. The language spoken at home was consistently Russian, and exposure to
English was minimal.

Tanja was born on 14-DEC-1993. The dates of the recordings, and Tanja’s age at
each recording, are as follows:

<table>
<thead>
<tr>
<th>File</th>
<th>Date</th>
<th>Age</th>
<th>File</th>
<th>Date</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tanja01</td>
<td>28-MAY-1996</td>
<td>2;5.14</td>
<td>Tanja09</td>
<td>20-SEP-1996</td>
<td>2;9.6</td>
</tr>
<tr>
<td>Tanja02</td>
<td>10-JUN-1996</td>
<td>2;5.27</td>
<td>Tanja10</td>
<td>25-OCT-1996</td>
<td>2;10.11</td>
</tr>
<tr>
<td>Tanja03</td>
<td>18-JUN-1996</td>
<td>2;6.4</td>
<td>Tanja11</td>
<td>08-NOV-1996</td>
<td>2;10.25</td>
</tr>
<tr>
<td>Tanja04</td>
<td>25-JUN-1996</td>
<td>2;6.11</td>
<td>Tanja12</td>
<td>11-NOV-1996</td>
<td>2;10.28</td>
</tr>
<tr>
<td>Tanja05</td>
<td>23-JUL-1996</td>
<td>2;7.9</td>
<td>Tanja13</td>
<td>15-NOV-1996</td>
<td>2;11.01</td>
</tr>
<tr>
<td>Tanja06</td>
<td>12-AUG-1996</td>
<td>2;7.29</td>
<td>Tanja14</td>
<td>22-NOV-1996</td>
<td>2;11.08</td>
</tr>
<tr>
<td>Tanja07</td>
<td>29-AUG-1996</td>
<td>2;8.15</td>
<td>Tanja15</td>
<td>04-DEC-1996</td>
<td>2;11.20</td>
</tr>
<tr>
<td>Tanja08</td>
<td>09-SEP-1996</td>
<td>2;8.26</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Publications using these data should cite:

7. Serbian

Darinka Andjelković
Nada Ševa
Jasmina Moskovljević

Eight children.
8. Slovenian – Žagar

The Slovenian corpus includes 37 transcriptions and video-files of child Slovenian that were obtained during the research project "Education and critical literacy constructions: perception, argumentation and interpretation" in 2005-2009, coordinated by dr. Igor Žagar and financed by the Ministry of Higher Education, Science and Technology RS. We would like to ask for the copies of articles that make use of our data. They should be sent to: igor.zagar@guest.arnes.si.

The recordings were primarily made for a study of argumentative patterns among children and not only to study linguistic patterns. Because many children were present at the recording site, there is often partial unintelligibility of the recordings (due to multiple overlaps of the sounds). Of all the recordings, it was important to transcribe the data that has been found relevant for the project (i.e. parts that include conversations with argumentation), some short parts that didn't include conversation (for example reading of the book) were not included. The data have been transcribed, and some parts of the transcriptions have been double-checked.

The aim of the recording was to assess some typical argumentation patterns in preschool children and in the child-directed speech. The transcription is done in colloquial written Slovenian with some phonological normalization used during the transcribing. The video material is digitalized and is available along with the transcription files.

Two kindergarten groups (ages 3-4 years and 5-6 years, normally developing population) were recorded for two weeks during their normal kindergarten activities, two recording hours per day, one in the morning and one in the early afternoon. Out of 40 recordings, 37 were transcribed. The children and the teachers speak Ljubljana dialect since they live there and go to kindergarten there, but there is no information available about the background of their parents.

Group 1 was recorded from 24-JAN-2005 to 04-FEB-2005 with two recordings each working day. Group 2 was recorded from 17-mar-2005 to 1-apr-2005 with two recordings every working day – one hour in the morning and one in the afternoon. The recordings were made in two rooms of the kindergarten Zelena jama, Ljubljana. The rooms are about 30 m2 large and fully equipped for approximately 20 children (toys, furniture, additional rooms for the bathrooms). The transcriber was Pavel Koltaj. We have the original consents from all the parents.

When using this corpus, please cite:


Group 1 Birthdates:
@Birth of 1: 10-OCT-2001
@Birth of 2: 13-AUG-2001
@Birth of 3: 17-AUG-2001
@Birth of 4: 19-SEP-2001
@Birth of 5: 26-NOV-2001
@Birth of 6: 04-AUG-2001
@Birth of 7: 21-SEP-2001
@Birth of 8: 02-JUL-2001
@Birth of 9: 22-NOV-2001
@Birth of 10: 06-AUG-2001
@Birth of 11: 22-JUN-2001
@Birth of 12: 21-APR-2001
@Birth of 13: 21-SEP-2001
@Birth of 14: 30-JAN-2002
@Birth of 15: 24-NOV-2001
@Birth of 16: 23-SEP-2001
@Birth of 17: 19-JUN-2001

Group 2 Birthdates:
@Birth of 1: 09-JUL-1999
@Birth of 2: 25-MAR-1999
@Birth of 3: 25-JAN-1999
@Birth of 4: 03-AUG-1999
@Birth of 5: 21-SEP-1999
@Birth of 6: 16-NOV-1999
@Birth of 7: 30-NOV-1999
@Birth of 8: 02-JUN-1999
@Birth of 9: 10-JUN-1999
@Birth of 10: 24-APR-1999
@Birth of 11: 10-AUG-1999
@Birth of 12: 29-SEP-1999
@Birth of 13: 25-OCT-1999
@Birth of 14: 18-MAR-1999
@Birth of 15: 20-MAY-1999
@Birth of 16: 09-JUL-1999
@Birth of 17: 23-JUN-1999