

# A Phonemic Corpus of Polish Child-Directed Speech

Luc Boruta and Justyna Jastrzebska

ALPAGE, Univ. Paris Diderot & INRIA

luc.boruta@inria.fr

Advances in **modeling language acquisition** are due not only to the development of machine-learning techniques, but also to the **increasing availability of data on child language** and child-adult interaction.

In the absence of high-quality recordings of child-directed speech or when models require such a representation for training data, **phonemically transcribed corpora** have commonly been used.

Models of psycholinguistic processes are expected to generalize to **typologically different (if not all) languages**. However, the best known corpora of child-directed speech have been developed for English or one of a small number of other languages. Polish is one of many **low-resource languages** when it comes to the evaluation of computational models of language acquisition.

We present a novel (and to our knowledge, the first) **phonemic corpus** of Polish child-directed speech.

## Sources of Data

We derived our phonemic corpus from the **Weist corpus** of Polish child-directed speech (Weist et al., 1984) that is freely available from the seminal CHILDES database (MacWhinney, 2000).

It contains 39 **CHAT**-formatted transcripts of non-elicited spontaneous verbal interactions involving four Polish-learning children (aged 1;7 to 2;6) and their respective caregivers.

For each utterance, the basic unit of data in the transcripts is an **orthographic transcription**, a **gloss**, and a **translation** to English.

## Phonemic Inventory


We used Jassem's (2003) description of Polish **phonemes** : /p, b, t, d, k, g, c, ʃ, m, n, ŋ, ŋ, f, v, s, z, ʃ, ʒ, ʒ, x, ts, dz, tʃ, dʒ, tɕ, dʒ, l, r, j, w/ for **consonants**, and /i, ɨ, u, e, o, a/ for **vowels**.

## Derived Corpus

The contributed corpus of Polish child-directed speech contains **15,364 utterances** (representing 11,194 types), **54,662 words** (5,712 types) and **225,324 phonemes** (37 types). This corpus is approximately twice the size of the now-standard **Brent/Ratner corpus** of English.

## Resources and License

All updated CHAT-formatted **transcripts** and the phonemic **lexicon** are distributed under the terms of the **Lesser General Public License for Linguistic Resources** (LGPL-LR).



\*EWA: to są dwie lalunie  
%pho: **tow sow dvje lalune**  
%eng: *these are two dolls*  
\*EWA: a co one robią  
%pho: **a tso one robiow**  
%eng: *what are they doing*  
\*EWA: oczka mają zamknięte  
%pho: **otʃka majow zamkɲente**  
%eng: *their eyes are closed*  
\*CHI: śpią  
%eng: *they are sleeping*  
%mor: V|sleep&IMPF:PRES:3P  
\*EWA: no zamknięte  
%pho: **no zamkɲente**  
%eng: *yes, closed*  
\*MOT: śpią laleczki  
%pho: **ɕpiow laletʃki**  
%eng: *they are asleep*  
\*MOT: śpią  
%pho: **ɕpiow**  
\*CHI: nie zamkną  
%eng: *they won't close (the eyes)*  
%mor: NEG|not V|close&PFV:FUT:3P  
\*MOT: nie zamkną oczek  
%pho: **ɲe zamknow otʃek**  
%eng: *will they not close their eyes*  
\*MOT: zamknęły bo śpią  
%pho: **zamkɲeɲwɨ bo ɕpiow**  
%eng: *they have closed them because they are asleep*  
\*CHI: już zamknęły  
%eng: *they have already closed (their eyes)*  
%mor: ADV|already V|close&PFV-PAST-NONVIR:3P  
\*MOT: już zamknęły  
%pho: **juʒ zamkɲeɲwɨ**  
%eng: *they have already closed (their eyes)*  
\*EWA: a jedno się otworzyło  
%pho: **a jedno ɕie otfoʒɨwo**  
%eng: *but one (eye) opened (by itself)*  
...