# A Phonemic Corpus of Polish Child-Directed Speech

**Luc Boruta**[1] **and Justyna Jastrzebska**[2]

[1]Univ. Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA, Paris, France
[1]École Doctorale Frontières du Vivant, Programme Liliane Bettencourt, Paris, France
[1]LSCP, Département d'Études Cognitives, École Normale Supérieure, Paris, France
[2]Univ. Paris Diderot, Sorbonne Paris Cité, UFR Linguistique, Paris, France
luc.boruta@inria.fr, justyna_jastrzebska@yahoo.fr

**Abstract**

Recent advances in modeling early 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 recordings of child-directed speech, or when models explicitly require such a representation for training data, phonemic transcriptions are commonly used as input data. We present a novel (and to our knowledge, the first) phonemic corpus of Polish child-directed speech. It is derived from the Weist corpus of Polish, freely available from the seminal CHILDES database. For the sake of reproducibility, and to exemplify the typical trade-off between ecological validity and sample size, we report all preprocessing operations and transcription guidelines. Contributed linguistic resources include updated CHAT-formatted transcripts with phonemic transcriptions in a novel phonology tier, as well as by-product data, such as a phonemic lexicon of Polish. All resources are distributed under the LGPL-LR license.

**Keywords:** child-directed speech; Polish; phonology

## 1. Introduction

Recent advances in modeling early 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 —throughout this study, we use *child-directed speech* as an umbrella term for *child-* and *infant-directed speech*, i.e. linguistic data children may use to bootstrap into language— or when models explicitly require such a representation for training data, phonemically transcribed child-directed corpora have commonly been used as input data.

Indeed, phonemic transcriptions have been used to develop and validate, among other tasks, computational models of the early acquisition of word segmentation, phonological knowledge, or both (Venkataraman, 2001; Peperkamp et al., 2006; Blanchard and Heinz, 2008; Daland and Pierrehumbert, 2010; Boruta et al., 2011, inter alios). Moreover (unless otherwise stated), computational models of psycholinguistic processes are expected to generalize to typologically different (if not all) languages (Gambell and Yang, 2004). Nonetheless, the best known corpora of child-directed speech have been developped for English or one of a small number of other languages, and Polish is one of many low-resource languages when it comes to the evaluation of computational models of language acquisition.

The purpose of this paper is to present a novel (and to our knowledge, the first) phonemic corpus of Polish child-directed speech that subsequent studies might use to determine how models of early language acquisition perform on Polish or, by extension, Slavic languages.

L. B. designed the study, analyzed the original corpus, and wrote the paper; J. J. developed the phonemic lexicon. Both authors discussed the results and implications at all stages.

## 2. Sources of Data

We derived our phonemic corpus from the Weist corpus of Polish child-directed speech (Weist et al., 1984; Weist and Witkowska-Stadnik, 1986) that is freely available from the seminal CHILDES database (MacWhinney, 2000). This corpus contains 39 CHAT-formatted transcripts of interactive, non-elicited, spontaneous verbal interactions involving four Polish-learning children (aged 1;7 to 2;6 at the time of recording) 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. Data was coded at the morphological level in the glosses, and no phonetic or phonemic information is available in a systematic manner. It is also worth noting that the audio recordings are freely available as WAV files from the Media section of the CHILDES database.

## 3. Deriving a Phonemic Corpus

The Brent/Ratner corpus of English child-directed speech (Brent and Cartwright, 1996) has now become the standard dataset for the evaluation of computational models of early language acquisition. Therefore, we followed the design and format of that corpus in order to derive our phonemic corpus of Polish from Weist et al.'s orthographic transcripts. For the sake of reproducibility and global consistency, most derivation steps were automated.

In keeping with usual practice, slashes $/\cdot/$ are used from this point forward to enclose phonemic transcriptions, and chevrons $\langle\cdot\rangle$ are used to enclose material from the CHAT-formatted orthographic transcripts.

### 3.1. Extracting Standard Child-Directed Utterances

As with the Brent/Ratner corpus, the first processing step consisted in the automatic extraction of the child-directed utterances from Weist et al.'s original transcripts, that is to

| | Bilabial | Labiodental | Postdental | Alveolar | Alveopalatal | Palatal | Velar |
|---|---|---|---|---|---|---|---|
| Nasal | m | | n | | ɲ | | ŋ |
| | m | | n | | N | | G |
| Plosive | p b | | t d | | | c ɟ | k ɡ |
| | p b | | t d | | | c J | k g |
| Fricative | | f v | s z | ʃ ʒ | ɕ ʑ | | x |
| | | f v | s z | S Z | 6 4 | | x |
| Affricate | | | ʦ ʣ | ʧ ʤ | ʨ ʥ | | |
| | | | T D | 1 2 | 7 5 | | |
| Lateral | | | l | | | | |
| | | | l | | | | |
| Flap/Trill | | | | r | | | |
| | | | | r | | | |
| Glide | | | | | | j | w |
| | | | | | | j | w |

Table 1: Consonants and glides of the phonemic inventory of Polish (IPA in serif, ASCII in monospaced typeface). Where symbols appear in pairs, the one to the right represents a voiced consonant.

say all utterances except the ones uttered by the so-called target child. Data for each of the four children were concatenated into a single meta-corpus.

Out of the raw 17,553 extracted utterances, only 15,364 (88%) complete and well-formed utterances were further selected to build the phonemic corpus. In order to control the trade-off between ecological validity and sample size, the following selection criteria were applied.

#### 3.1.1. Exclusion Criteria

First, utterances containing actions without speech, unidentifiable, guessed or untranscribed material (annotated with the ⟨0⟩, ⟨xxx⟩, ⟨[?]⟩, and ⟨www⟩ CHAT markers, respectively) were automatically discarded from the corpus as, undeniably, no proper phonemic transcription may be recovered from those utterances. Following the same argument, incomplete utterances, or utterances containing phonological fragments (annotated with the ⟨&⟩ marker) or incomplete words were also discarded.

Finally, utterances containing non-standard words such as babbling, onomatopoeia, interjections, word plays, neologisms, child-invented, second-language or family-specific forms (transcribed with the ⟨@b⟩, ⟨@o⟩, ⟨@i⟩, ⟨@wp⟩, ⟨@n⟩, ⟨@c⟩, ⟨@s⟩, and ⟨@f⟩ markers, respectively) were discarded as their transcription and/or their frequency of occurrence might prove problematic.

#### 3.1.2. Deletion Criteria

Conversely, as they do not affect the integrity of the utterances, other CHAT-formatted annotations such as paralinguistic material, pause symbols, punctuation marks, utterance terminators and utterance linkers were merely deleted from the orthographic transcriptions.

An example of this is the trailing-off terminator ⟨+...⟩ which marks the end of an incomplete (but not interrupted) utterance. By deleting this marker in, for example, ⟨jeszcze nie +...⟩ (*not yet*), the utterance would then be mapped to the phonemic transcription /jeʃʧe ɲe/.

It is also worth mentioning that, when annotated as single phrases in the original transcripts with the repetition marker ⟨[x $n$]⟩, phrasal repetitions were not expanded; for example, the utterance ⟨dzień dobry, dzień dobry, dzień dobry⟩ (*good morning, good morning, good morning*) would only be phonemically transcribed as /ʥieɲ dobrɨ/ if coded as ⟨dzień dobry [x 3]⟩. Hence, studying repetitions or disfluencies, or computing corpus statistics (such as the ubiquitous mean length of utterance measure, a.k.a. MLU) from this derived representation of the data might result in observations far different from those obtained using CHAT-specific software such as CLAN (MacWhinney, 2000).

Deleting punctuation marks was done to create a stripped-down representation of the data, identical to the one used in the Brent/Ratner corpus: one utterance per line, containing only space-separated, phonemically-transcribed words with a one-to-one mapping between the phonemes and the symbols used to represent them.

### 3.2. Extracting the Lexicon

Once the proper subset of complete, well-formed utterances was selected, the following step consisted in the automatic extraction the attested lexicon from the candidate corpus, i.e. the set of all occurring orthographic words. As in the Brent/Ratner corpus of English, word transcriptions were designed so that each orthographic word form is matched with a single phonemic transcription.

As grapheme-to-phoneme relations in Polish are regular (though complex), phonemic transcriptions could have been obtained automatically (Steffen-Batogowa, 1975; Jassem, 2003). However, because of the relatively small size of the lexicon in the corpus at hand (5,712 types), we sacrificed some reproducibility for the sake of quality, and transcribed the lexicon manually.

### 3.3. Transcription Scheme

As for the definition of the phonemic inventory, we used Jassem's authoritative description of Polish phonemes (Jassem, 2003). The resulting phonemic inventory of Polish is presented in Table 1 for consonants and glides, and in Figure 1 for vowels.

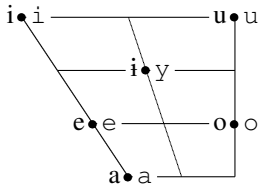


Figure 1: Vowels of the phonemic inventory of Polish (IPA in serif, ASCII in monospaced typeface).

Because the implementation of many computational models of early language acquisition —legacy, such as Venkataraman's NGS-u (Venkataraman, 2001), or not, such as Goldwater et al.'s DP (Goldwater et al., 2009)— do not accomodate Unicode characters denoting the symbols of the international phonetic alphabet (henceforth IPA), each phoneme was mapped to an ASCII character in the released resources.

## 4. Derived Corpus

Finally, the whole phonemic corpus was automatically reconstructed replacing, in each extracted child-directed utterance, every orthographic word form by its phonemic transcription. The final derived phonemic corpus of Polish child-directed speech contains 15,364 utterance tokens (representing 11,194 types), 54,662 words (5,712 types) and 225,324 phonemes (37 types). Moreover, it is worth noting that this corpus of Polish is approximately twice the size of the Brent/Ratner corpus of English: it contains 1.6 times more utterances (in terms of tokens), 1.6 times more words, and 2.4 times more phonemes. Phoneme frequencies observed in the corpus are reported in Table 2.

Aside from the stripped-down format used in the Brent/Ratner corpus, and because subsequent studies might require to enforce ecological validity, e.g. by extracting utterances addressed to a single child, the resulting phonemic transcriptions of the Weist corpus were also included as a novel phonology tier (denoted `%pho`) in updated CHAT-formatted transcripts. An excerpt of such a transcript is presented in Figure 2.

| Phoneme | Frequency |
|---|---|
| a a | 23586 |
| e e | 23021 |
| o o | 21632 |
| t t | 14422 |
| i i | 13746 |
| k k | 10062 |
| j j | 8596 |
| w w | 8197 |
| u u | 7420 |
| m m | 7072 |
| ɨ y | 6485 |
| v v | 5966 |
| n n | 5934 |
| p p | 5743 |
| r r | 5614 |
| b b | 5114 |
| s s | 5019 |
| ʃ S | 4859 |
| ɕ 6 | 4577 |
| ɲ N | 4215 |
| d d | 3994 |
| ʦ T | 3728 |
| l l | 3171 |
| z z | 3112 |
| ʨ 7 | 3083 |
| ʧ 1 | 2825 |
| ʒ Z | 2623 |
| ʥ 5 | 2561 |
| ɡ g | 2464 |
| x x | 1824 |
| f f | 1336 |
| ŋ G | 985 |
| ʤ D | 585 |
| ɟ J | 391 |
| ʑ 4 | 326 |
| c c | 41 |
| ʣ 2 | 26 |

Table 2: Phoneme frequencies in the derived corpus (IPA in serif, ASCII in monospaced typeface).

## 5. On the Usefulness of such Corpora

To further emphasize the need for computational models of early language acquisition to be evaluated using typologically different languages, we compare the performance of a well-known unsupervised model of the acquisition of word segmentation (Venkataraman, 2001) on three phonemic corpora (derived) from the CHILDES database: the aforementioned Brent/Ratner corpus of English, the Johnson/Demuth corpus of Sesotho (Johnson, 2008), and our novel corpus of Polish.

Providing a thorough analysis of the discrepancies between this model's performance on English, Sesotho and Polish is beyond the scope of this paper. Yet, a standard quantitative evaluation gives enough weight to the argument at hand. As Venkataraman's model is incremental, its output is conditional on the order in which input utterances are processed; we thus report average segmentation F-scores (together with standard deviations within parentheses) observed over 100 distinct random permutations of the input corpora. The results are categorical: the F-score of the word segmentation model is 66.8% (2.9) for English, 50.4% (3.3) for Polish, and only 28.4% (7.1) for Sesotho.

As evidenced here, evaluating the performance of a computational model of early language acquisition using data from only one language offers no guarantee whatsoever as to the performance of the model on other languages, especially if typologically unrelated; hence the need for a special effort in creating appropriate linguistic resources. To our knowledge, the novel phonemic corpus of Polish child-directed speech we presented in this paper is the first resource of this kind made available for Polish and, by extension, Slavic languages.

```
*CHI:   radio ma pani .
%eng:   the lady has a radio .
%xmor:  N|radio-NEUT:ACC:S V|have&IMPF:PRES:3S N|lady-FEM:NOM:S .
*MOT:   radio ma pani .
%pho:   radjo ma paɲ
%sit:   Marta's room is full of various toys . there is also a pram with two dolls .
%eng:   the lady has a radio .
*MOT:   oczywiście .
%pho:   otʃivi͡ɕt͡ɕie
%eng:   of course .
*EWA:   to są dwie lalunie .
%pho:   tow sow dvje laluɲe
%eng:   these are two dolls .
*EWA:   a co one robią ?
%pho:   a t͡so one robiow
%eng:   what are they doing ?
*EWA:   oczka mają zamknięte .
%pho:   otʃka majow zamknjeɲte
%eng:   their eyes are closed .
*CHI:   śpią .
%eng:   they are sleeping .
%xmor:  V|sleep&IMPF:PRES:3P .
```

Figure 2: Beginning of the updated `marta1.cha` transcript. Added phonemic transcriptions are in the bold `%pho` tier.

## 6. Resources and License

All aforementioned resources, derived (the updated CHAT-formatted transcripts) or original (the phonemic lexicon), are distributed under the terms of the Lesser General Public License for Linguistic Resources (LGPL-LR).
In addition to being included into LREC's LRE Map, these resources were also recontributed to the Derived Corpora and Counts section of the CHILDES database.

## 7. Acknowledgements

This work was supported in part by a graduate fellowship from the French Ministry of Higher Education and Research granted to the first author.

## 8. References

D. Blanchard and J. Heinz. 2008. Improving word segmentation by simultaneously learning phonotactics. In *Proceedings of the Twelfth Conference on Natural Language Learning*, pages 65–72.

L. Boruta, S. Peperkamp, B. Crabbé, and E. Dupoux. 2011. Testing the robustness of online word segmentation: effects of linguistic diversity and phonetic variation. In *Proceedings of the Second Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–9.

M. R. Brent and T. A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125. Data available online from http://childes.psy.cmu.edu/derived/brent_ratner.zip.

R. Daland and J. B. Pierrehumbert. 2010. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

T. Gambell and C. Yang. 2004. Statistics learning and universal grammar: modeling word segmentation. In *Proceedings of the Twentieth International Conference on Computational Linguistics*.

S. Goldwater, T. L. Griffiths, and M. Johnson. 2009. A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, 112(1):21–54.

W. Jassem. 2003. Illustrations of the IPA: Polish. *Journal of the International Phonetic Association*, 33(1):103–107.

M. Johnson. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27. Data available online from http://childes.psy.cmu.edu/derived/sesotho.zip.

B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Elbraum Associates.

S. Peperkamp, R. Le Calvez, J. P. Nadal, and E. Dupoux. 2006. The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41.

M. Steffen-Batogowa. 1975. *Automatyzacja transkrypcji fonematycznej tekstów polskich [Automatic phonemic transcription of Polish texts]*. Państwowe Wydawnictwo Naukowe.

A. Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372.

R. M. Weist and K. Witkowska-Stadnik. 1986. Basic relations in child language and the word order myth. *International Journal of Psychology*, 21:363–381.

R. M. Weist, H. Wysocka, K. Witkowska-Stadnik, E. Buczowska, and E. Konieczna. 1984. The defective tense hypothesis: on the emergence of tense and aspect in child Polish. *Journal of Child Language*, 11:347–374. Data available online from http://childes.psy.cmu.edu/data/Slavic/Polish/Weist.zip.