The Polish Frequency List of Child Directed Speech in Comparison to the Standard Polish Language

Magdalena ŁUNIEWSKA[1], Jakub SZEWCZYK[2], Bartłomiej ETENKOWSKI[3], Ewa HAMAN[4]
[1]University of Warsaw, Poland
[2]Jagellonian University in Cracow, Poland
[3]Charles University in Prague, Czech Republic
luniewskam@gmail.com

Child Directed Speech (CDS) is claimed to play a crucial role in language acquisition (Clark, 1993; 2009). In particular, word frequency in CDS has an influence on word learning order and pace (Dale, Goodman & Li, 2008; Bannard & Matthews, 2008). Word frequency in CDS is a language specific factor, making it necessary to individually measure frequency for each language. Nowadays it is obvious that Child Directed Speech differs from standard (adult directed) speech (e.g. Fernald *et al.*,1989; Tomasello *et al.*, 1990) but it is still unclear in which particular areas they diverge in Polish. In this report we describe the Polish Frequency List of Child Directed Speech and compare it to the standard Polish language (both spoken and written).

Since Polish is a highly inflected language, a single lexeme can have more than thirty different forms. The major benefit of the Polish Frequency List of CDS (which contains not only inflected forms of the words as they appear in the corpora) is that it is also lemmatized and contains summed frequencies for the base forms (lexemes). It is also almost free of typographical errors which makes it more reliable and easier to be used for quantitative analyses of CDS, designing psycholinguistic experiments, the preparation of materials for language disorders diagnosis, therapy and countless other projects.

The List is based on seven corpora, some of them already available from the CHILDES (Szuman corpus and Weist corpus, MacWhinney, 2000), and the others to be added in the future. Another important source is the Polish Children's Speech Corpus (http://www.kognitywistyka.amu.edu.pl/js/korpus.html). All corpora used include more than 1,179,000 word tokens (with more than 794,000 word tokens of speech directed to children aged between 0;10 and 7;0), about 44,000 word types, and 21,000 different lexemes.

In the first part of the report we describe in detail the characteristics of the material which the child directed and children speech corpus is based on (circumstances of recordings, the characteristics of speakers) and the way the Polish Frequency List of CDS was prepared.

In the second part of this report the Polish Frequency List of CDS is compared to the standard spoken Polish frequency list (Conversational Spoken Corpus of Polish), and to various thematic subcorpora of the National Corpus of Polish. We compare frequency distribution of a representative set of words in all these corpora, and their similarity index, in order to reveal characteristics of the CDS corpus. We also compare some grammatical characteristics of the corpora, such as frequencies of grammatical categories (case, tense, person).

References

Bannard, C. & Matthews, D. (2008). 'Stored Word Sequences in Language Learning. The Effect of Familiarity on Children's Repetition of Four-Word Combinations.' *Psychological Science* 19: 241-248.

Clark, E. V. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Clark, E. V. (2003). *First Language Acquisition*. Cambridge: Cambridge University Press.

Dale, P.S., Goodman, J.C. & Li, P. (2008) ‘Does frequency count? Parental input and the acquisition of vocabulary.’ *Journal of Child Language* 35: 515-531.

Fernald, A., Taeschner, T., Dunne, J., Papousek, M., de Boysson-Bardies, B. & Fukui, I. (1989). ‘A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants.’ *Journal of Child Language* 16: 477-501.

McWhinney, B. & Snow C. (1985). ‘The Child Language Data Exchange System.’ *Journal of Child Language* 12: 271-296.

Tomasello, M., Conti-Ramsden, G. & Ewert, B. (1990). ‘Young children's conversations with their mothers and fathers: Differences in breakdown and repair.’ *Journal of Child Language* 17: 115-130.