Dealing with Bad Data in Language Acquisition

Paula Fikkert
Centre for Language Studies, Radboud University Nijmegen

The field of language acquisition is flourishing and many important insights have been gained based on production and perception studies on language acquisition. We now know that language acquisition starts long before children utter their first word, that infants tune into their native language somewhere during the first year of life, that children’s first utterances share characteristics in many languages studied to date, etc.

Yet, there are also reasons to reflect on the nature of the acquisition data and the insights they reveal in both language perception and production, albeit for different reasons.

Language acquisition research started with handwritten transcriptions of children’s utterances based on parental observation, but the data collection improved drastically in the last decades among others due to technological improvements. It is now easy to collect large amounts of data during whole day recordings and record everything a child says and hears. In addition, the CHILDES initiative has done an excellent job making data available. It has been extended to make phonetically transcribed data available often with the original soundfiles or movies (ChildPhon). Still, good data are hard to get and ‘bad’ data abundant.

Since the 70’s of last century perception studies have informed language acquisition research in addition to production studies, and nowadays baby research centres with eye trackers and head turn boots are appearing everywhere. Methods for collecting perception data have improved, also aided by technological advances. The numerical data (looking times, reaction times, etc.) are often interpreted to reveal insights into linguistic knowledge of the participants, but it is important to remain critical.

In both cases the data often suffer from all problems mentioned in the call for this workshop: the data are always noisy and variable, often incomplete, one-sided (or limited in scope) and not seldom conflicting. In perception studies there is also a high number of studies with null results adding to the problem.

In this talk I’ll present (a) the types of data that are used in language acquisition research; (b) discuss the problems with these data, and (c) current trends and initiatives to deal with those problems.