Detecting chunks in child bilingual code-mixing – the role of input

Nikolas Koch, LMU Munich, koch@daf.lmu.de
Antje Endesfelder Quick, Leipzig University, antje.quick@uni-leipzig.de
Stefan Hartmann, HHU Düsseldorf, hartmast@hhu.de

Instances of code-mixing, i.e. utterances that combine elements from two or more languages, in language acquisition are often regarded as particularly creative utterances of children that cannot be attributed to the respective input (cf. De Houwer 2009: 44). In recent years these utterances have seen increased interest from a usage-based perspective, in which patterns of language use play a pivotal role. Usage-based approaches have developed a number of methods that allow for detecting patterns from naturalistic data. So far, these methods have mostly been applied to monolingual data. In this talk, we evaluate one of these methods with regard to its performance when applied to code-mixing (CM) data: the so-called Chunk-based-Learner (CBL) model (McCauley & Christiansen 2017, 2019). This method focuses on chunking processes and makes it possible to automatically detect patterns (chunks) in speech data by combining simple frequency measures and basic learning mechanisms, such as entrenchment. In order to detect chunks the model uses a simple metric, backward transitional probabilities (BTP). The BTP value indicates how likely it is that the current word is preceded by the word that precedes it in the current context. To identify chunks, the model constantly calculates the average BTP and if the current BTP value is above the average BTP, the word pair is regarded as a potential chunk. If the BTP of the word pair falls below the current average BTP, a boundary between the two words is assumed.

In this study, we want to apply the CBL to code-mixed utterances (*n*= 3492) of a German-English bilingual child, Fion, aged 2;3 to 3;11 and to his input data (*n*= 228,221). First results show that CM utterances exhibit a greater number of chunks compared to the child’s monolingual utterances. Secondly, we also found that the patterns in the CM utterances can largely be attributed to chunks that occur in the input. Here, about two-thirds (3,080 utterances) of the language boundaries within CM utterances coincide fully ([ein kleinen][shak]; 'a little shark') or partially ([I have][that werkzeug]; 'I have the tool') with chunk boundaries. In 426 CM utterances, there is no overlap between speech and chunk boundaries. These are often bilingual chunks such as [time out machen] ('take a timeout') or [ein anderer frog] ('another frog') which have no occurrences in the parental input but might be consolidated by the brother, who is a bilingual child himself. The question arises to what extent the extracted chunks in the boy's CM utterances can be attributed to input from his parents, his brother, or to his own utterances. Especially in the case of the bilingual chunks, we expect them to be consolidated not by his caregivers but by utterances of his brother or himself. Furthermore, we will test the hypothesis that the chunks that adhere to language boundaries can be attributed primarily to input from caregivers. In this context, the question arises whether these are also the chunks that occur frequently in the boy's monolingual utterances. The results and advantages and disadvantages of the method are discussed in the light of a usage-based approach.

References

De Houwer, Annick. 2009. *Bilingual First Language Acquisition*. Bristol: Multilingual Matters.

McCauley, Stewart M. & Morten H. Christiansen. 2017. Computational Investigations of Multiword Chunks in Language Learning. *Topics in Cognitive Science* 9(3). 637–652.

McCauley, Stewart M. & Morten H. Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological Review* 126(1). 1–51.