Contribution ID: **60** Type: **not specified**

Co-occurrence statistics as a language-dependant cue for speech segmentation

Distributional regularities in spoken languages are informative about speech units (words), such that the dependencies are stronger within words than between words; this could therefore be a useful cue for word learning even without rich language-specific knowledge. When tested with artificial speech that contains distributional regularities but no other cues for word boundaries, infants and adults show the ability to extract words from it.

However, artificial languages used in laboratories are usually much more regular than spoken languages, and a more complex strategy might be necessary to segment speech in a natural language than in an artificial stream. Various models of speech segmentation have been proposed in the literature to solve this issue, using different spoken corpora as their input.

Additionally, it is not clear whether the information that can be extracted from distributional regularities is comparable across different languages, such that an uninformed learner could use the same strategy regardless of the input language. To explore this latter question, we model two learning strategies based on transitional probabilities using child-directed speech corpora from nine languages. We show that languages vary as to which statistical segmentation strategies are most successful. The variability of the results can be partially explained by systematic differences between languages, such as rhythmical differences. This in turn indicates that infants may have to primarily rely on non-statistical cues when they begin their process of speech segmentation.

Primary author: Dr SAKSIDA, Amanda (SISSA)

Co-authors: Dr LANGUS, Alan (SISSA); NESPOR, Marina (SISSA)

Presenter: Dr SAKSIDA, Amanda (SISSA)