Charles Yang, University of Philadelphia - How Children Overcome the Sparsity Problem

The LREC community needs no reminder of the sparse data problem. There just isn't enough data for complex statistical models of language, nor does the corpus always contain data of critical theoretical interest. Against this background, it is remarkable that young children (almost) flawlessly learn the essential ingredients of their native language, all in an unsupervised setting and with as little as 10 million words of very simple speech.

I will discuss two central issues in the study of children language acquisition to draw out useful connections to corpus and computational linguistics. The first is a statistical problem. How do we determine the underlying system given a corpus of language data? Is the data generated by a productive rule or by the storage and retrieval of lexically specific items such as collocations? I present a statistically rigorous solution along with evidence that even very young children's language is rule based, contrary to claims in the usage-based literature. The second problem is learning theoretic. Given that children learn systematic rules very early on, how do they do it--in the face of sparse data and having to contend with rules often laden with exceptions? I introduce a simple principle of rule learning that is well supported by corpus, behavioral, and empirical studies of child language. The principle has the mathematical property that rules are easier to learn if the learner has smaller vocabularies, which allows us to draw the seemingly paradoxical conclusion that data sparsity does not hinder but in fact enable the successful acquisition of language.