# Computational modeling of language acquisition

Rishabh Nigam , Shubhdeep Kochar
Mentor : Prof. Amitabh Mukherjee
Department of Computer Science And Engeneering
IIT Kanpur , India

## Problem :

Develop a computer model of language acquisition in the form of a generative grammar that is algorithmically learnable from realistic corpus data , viable in its large scale quantitative performance

## Introduction and literature:

Language Acquisition is the process by which humans acquire the capacity to perceive and comprehend language, as well as to produce and use words to communicate . In this project we try to develop and implement algorithms which learn from realistic corpus data .

For measuring the generative performance of the grammar we use two parameters Recall and Precision .Recall is defined as the proportion of unfamiliar sentences that a parser based on the grammar accepts .Precision is defined as the proportion of novel sentences generated by the grammar that are deemed acceptable by native-speaker subjects, preferably in a blind, controlled test . It is trivially easy to construct a grammar which has a precision of 1 or a grammar with recall 1, but recall of the former grammar and the precision of the latter will be very poor. What we aim at is to have high values for both recall and precision.

## Approach :

There have been a couple of algorithms to tackle this problem :
*ADIOS algorithm* : The ADIOS algorithm rests on two principles : (1) probabilistic inference of pattern significance and (2) recursive construction of complex patterns. ADIOS starts by representing a corpus of sentences as an initially
highly redundant directed graph, in which the vertices are the lexicon entries and the paths correspond to corpus sentences.
*Context algorithm* : In ConText, the distributional statistics of a word or a sequence of words (w) are determined by the surrounding words (i.e. local context). Sequences that are closer than D (a user-defined parameter) are viewed as equivalent and are clustered together. At the end of the clustering procedure, sequences belonging to the same cluster are assumed to be substitutable or equivalent. This way we generate different sentences based on the corpus .

## Why this problem :

In psycholinguistics, the main challenge is to discover the nature of grammar –the knowledge of language as it is represented in the brain . Much of the effort in the computational modeling of language development focuses on understanding specific phenomena, such as word segmentation . More general and abstract work typically involves computational models capable of learning certain classes of formal languages generated by small artificial grammars . In this project what we try to do is

to implement algorithm capable of learning a grammar that is generative of the target language, given a realistic corpus of child-directed speech . This way we will be studying language acquisition by infants of age less than one . The process of language acquisition can be used to acquire language by the bots in the real world making communication between the bots and human better . This is still a very open field where a lot of work can be done .

## Database :

The database that we shall be using is the CHILDES (Child Language Data Exchange System) database . This is an open database which can be downloaded at http://childes.psy.cmu.edu/data/Eng-USA/

## References :

[1] Heider. Waterfall ,Ben Sandbank,Luca Onnis and Shimon Edelman , An empirical generative framework for computational modeling of language acquisition* : Cambridge University Press 2010 **http://kybele.psych.cornell.edu/~edelman/Waterfall-Sandbank-Onnis-Edelman-JCL10.pdf**