Natural Language Processing and Early Language Acquisition

Supervisor:
Thomas Hills, Department of Psychology (contact: t.t.hills@warwick.ac.uk)

Overview

This project would apply recent developments in natural language processing and statistical mechanics to investigate the problem of early language acquisition in children.

Background:

How children acquire language is a longstanding problem in linguistics and human psychology. Recent developments in large-scale statistical approaches to text corpora, and network analyses of early learned words, provide powerful new tools for understanding the structure of learned information and how this is related to the statistical structure of speech directed to children. This project could involve one of two projects:

1) Applying statistical methods to corpora of early language spoken to children, for which we currently have over 4 million words in the CHILDES corpus. Adult-directed speech corpora would also be used for comparison. This work would look beyond frequency to sequential order statistics, such as bursts and lulls of word usage or associative representations. This work is fairly clean, as the tools to address these problems have been developed for other corpora and have yet to be applied to child-directed speech.
2) Bilingual acquisition: We currently have longitudinal data for more than 600 children's first learned words, between the ages of 1.5 and 3 years. These words will be combined into networks for monolingual and bilingual children and the statistical structure of these networks will be investigated to understand how word acquisition influences subsequent word acquisition.

Prerequisites and Future Prospects

Students interested in this project should have a command of computer programming and be interested in learning about network analysis and statistical analyses of large data sets. This work has broad implications for understanding language acquisition in both typical and atypical children and could lead to future (Ph.D.) work on language acquisition and natural language processing.

References:

Altmann, E. G., Pierrhumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: bursts, lulls, and scaling in the temporal distribution of words. *PLoS* One, 4, e7678.

Hills, T. (2012). The company that words keep: Comparing child and adult-directed language. *Journal of Child Language*, 1-19. Available online at doi:10.1017/S0305000912000165.

Serrano, M, Flammini, A., & Menczer, F. (2009). Modeling statistical properties of written test. *PLoS One*, 4, e5372.