Information Theory and Conceptual Physics

Supervisor: Thomas Hills (Psychology)

The aim of this project is to use tools from statistical physics and information theory to understand the statistical properties of concepts in language. Concepts can be derived from the statistical properties of millions of words in large natural corpora (e.g., Google Ngrams) using semantic space models (e.g., LSA). Conceptual properties can then be investigated for statistical laws such as Zipf's law, properties of new concepts as they enter the language, and properties of conceptual distributions as they change over time. The principle new idea is applying tools from statistical physics to concepts (distributional properties of individual words) instead of word frequency, the latter having studied fairly extensively.

This work can lead to PhD work and is useful for many applied projects in historical language analysis (see references).

References:

Hills, T., Proto, E., & Sgroi, D. (2015) Historical analysis of national subjective wellbeing using millions of digitized books. *IZA Discussion Paper No. 9195.*

Hills, T., & Adelman, J. (2015). Recent evolution in the learnability of American English from 1800 to 2000. *Cognition, 143,* 87-92.

Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific reports*, *2*.