Jona Sassenhagen
(Goethe-Universität, Frankfurt)

Triangulating Meaning between Philosophy, Artificial Intelligence and Cognitive Neuroscience

Empiricist theories of meaning (e.g., Churchland, 1992; Prinz, 2005) entail that concepts can be decomposed and learned. In contrast, Jerry Fodor (e.g., Fodor & LePore, 1999; Fodor, 2008) argues that no concept can learned, because inductively acquiring a concept requires having it available for a hypothesis test - i.e., concept learning is circular. Recent advances in Machine Learning/Artificial Intelligence (Mikolov et al., 2012) indicate that distributional learning can result in word representations with high applicability for computational linguistic tasks. I demonstrate that these representations are surprisingly resistant to Fodor's Circularity Critique - making them a *potential* candidate for an architecture of human conceptual knowledge. But how closely do they resemble our *actual* cognitive architecture? To probe this, distributionally learned representations were encoded in brain activity, and these encodings compared against real brain activity. In addition, orthographic and intermediate representations were encoded. Results from a series of experiments establish distributionally learned representations as a strong contender for our knowledge of the meaning of words. They also support an interactive, distributed coding of word meaning (rather than a strictly modular architecture). Methodologically, the techniques developed to test these questions can be leveraged to answer wide-ranging questions about the neurocognitive nature of the representation of linguistic knowledge beyond experimental paradigm-oriented research.