

## Living Language and Embodied Cognition: Thoughts on the Sciences of Language Inspired by the LLM Challenge

Chu-Ren Huang, The Hong Kong Polytechnic University

A linguistic act, in its simplest and most essential form, is the exchange of information between two people (i.e., two intelligent agents), with two design characteristics: first, the shared access to the expressed linguistic content and second, the lack of access to each other's internal cognitive processes. I argue in this talk that current linguistic theories are not designed to account for these two design characteristics of language. I further demonstrate that, by accounting for these two design features, language sciences can leverage large language models (LLMs) to advance our understanding of language, cognition, and how humans interact with the environment.

In order to set the stage for this argument, I first introduce the Russian doll (or nested self-hypernym) metaphor for the lexical conceptualization of 'language' shared by most (if not all) languages in the world. We use the lexical concept of 'language' to refer to a full range of different conceptually related entities, e.g. Chinese as both the modern language we speak and the inherited language that defines our culture, and language as the universal ability to acquire the system (the capitalized Language) to the choices of specific words (e.g., 'watch your language'). Second, I observe that sensory impairments do not preclude full development of linguistic and cognitive competence. These two observations, along with evidence that LLMs can learn shared sensorimotor understandings from collective data, suggest we need to take a closer look at the role of language data and how it is structured in cognition.

I then argue that, contrary to commonly held assumptions, the dynamic collection of language data is a constant in human cognition, anchoring the diversity of individual experiences and the variations of human brains. The power of language lies in how it enables the effective learning of knowledge grounded on shared language data. LLMs function by applying massive computing power to an unprecedented collection of language big data. Humans thrive by leveraging our embodied experiences and the embodied architecture of our cognition to learn from a smaller set of shared data. This data-centered view provides a straightforward account of language as a self-adoptive complex system, resulting from the dynamic growth of shared data/experience. In addition, the nested self-hypernyms conceptualization of 'language' captures the essence of language as a complex system while allowing the foregrounding of different levels of its sub-systems in daily human life.

This new perspective points to a different approach to the multi-brain alignment dilemma. That is, we understand each other not by accessing or guessing at each other's brain activities but by deriving shared information from our shared linguistic data. Thus, I suggest that the most unique characteristic of human language is its embodied architecture and its instantiation as an ever-evolving collection of shared language data. The information from this shared data architecture has paved a path from LLMs to AI, but it is its embodied architecture that remains a key distinction between human intelligence and machine intelligence.