Commentary on Michael A. Arbib/From monkey-like action recognition to human
language: An evolutionary framework for neurolinguistics

Language Evolution: Body of Evidence?

Chen Yu
Department of Computer Science
University of Rochester
Rochester, 14627
U.S.A.
585-275-1443
yu@cs.rochester.edu
http://www.cs.rochester.edu/~yu/

Dana H. Ballard
Department of Computer Science
University of Rochester
Rochester, 14627
U.S.A.
dana@cs.rochester.edu
http://www.cs.rochester.edu/~dana/

Abstract

Our computational studies of infant language learning estimate the inherent
difficulty of Arbib's proposal. We show body language provides a strikingly helpful
scaffold for learning language that may be necessary but not sufficient owing to
the absence of sophisticated language in other species. The extraordinary language
abilities of Homo sapiens must have evolved from other pressures, such as sexual
selection.

Arbib's article provides a complete framework of how humans but not monkeys have
language-ready brains. A centerpiece in hominid language evolution is based on the
recognition and production of body movements, particularly hand movements, and
their explicit representation in the brain, termed the mirror property.

How can we evaluate this proposal? One way is to take a look at infant language
learning. The infant has evolved to be language ready, but nonetheless, examining

the steps to competency in detail can shed light on the constraints that evolution had to deal with. In a similar manner to language evolution, the speaker (language teacher) and the listener (language learner) need to share the meanings of words in a language during language acquisition. A central issue in human word learning is the mapping problem—how to discover correct word-meaning pairs from multiple co-occurrences between words and things in an environment, which is termed reference uncertainty by Quine (1960). Our work in Yu, Ballard & Aslin (submitted) and Yu & Ballard (2004) shows that body movements play a crucial role in addressing the word-to-world mapping problem and the body's momentary disposition in space can be used to infer referential intentions in speech. By testing human subjects and comparing their performances in different learning conditions, we find that inference of speakers' intentions from their body movements, which we term embodied intentions, facilitate both word discovery and word-meaning association. In light of these empirical findings, we have developed a computational model that can identify the sound patterns of individual words from continuous speech using non-linguistic contextual information and employ body movements as deictic references to discover word-meaning associations. As a complimentary study in language learning, we argue that one pivotal function of a language-ready brain is to utilize temporal correlations among language, perception and action to bootstrap early word learning. Although language evolution and language acquisition are usually treated as different topics, the consistence of the findings from both Arbib's work and our work does show a strong link between body and language. Moreover, it suggests that the discoveries in language evolution and those in language acquisition can potentially provide some insightful thoughts to each other.

Language (even protolanguage) is about symbols and those symbols must be grounded so that they can be used to refer to a class of objects, actions, or events. To tackle the evolutionary problem of the origins of language, Arbib argues that language readiness evolved as a multimodal system and supported intended communication. Our work confirms Arbib's hypothesis and shows that a language-ready brain is able to learn words by utilizing temporal synchrony between speech and referential body movements to infer referents in speech, which leads us to ask an intriguing question -- how the mirror system proposed by Arbib can provide a neurological basis for a language learner to use body cues in language learning?

Our studies showed quantitatively how body cues that signal intention could aid infant language learning. Such intentional body movements with accompanying visual information provided a natural learning environment for infants to facilitate linguistic processing. Audio, visual and body movement data were collected simultaneously. The non-speech inputs of the learning system consisted of visual data, head and hand positions in concert with gaze-in-head data. The possible meanings of spoken words were encoded in this non-linguistic context, and the goal was to extract those meanings from raw sensory inputs. Our method first utilized eye and head movements as cues to estimate the speaker's focus of attention. At every attentional point in time, eye gaze was used as deictic reference (Ballard, Hayhoe, Pook, & Rao 1997) to find the attentional object from all the objects in a scene and each object was represented by a perceptual feature consisting of color, texture and shape features. As a result, we obtained a temporal sequence of

possible referents. Next, a partitioning mechanism categorized spoken utterances represented by phoneme sequences into several meaning bins, and an expectation-maximization algorithm is employed to find the reliable associations of spoken words and their perceptually grounded meanings. Detailed descriptions of machine learning techniques can be obtained from Yu and Ballard (2004). The learning result is that this system can learn over 85% of the correct word-meaning associations correctly given that the word has been segmented. Considering that the system processes raw sensory data, and our learning method works in unsupervised mode without manually encoding any linguistic information, this level of performance is impressive.

Such results are very consistent with Arbib's proposal that these body constraints served to start language development on an evolutionary scale. However it leaves unanswered question of why Homo sapiens. Arbib's argument seems to be that if a plausible sequence of steps is laid out, and the "height" or difficulty transiting each step is small then somehow evolution should have been compelled to follow this path. But our sequence of steps in the model of infant language learning also has small steps - recognize body movements, recognize intentions as communicated with body movements, recognize attentional objects in a scene, recognize the sounds that accompany these movements. These steps would be accessible by a variety of social species and yet they were only traversed by us. Arbib makes special use of the hand representations suggesting that perhaps humans had an edge in this category provided the needed leverage. It is again very plausible, yet our studies show that you can get quite far by just hanging sounds on the end of the eye fixations and hand movements. From our point of view, any animal species that could communicate intention through body movement had the possibility of developing some kind of language. Thus it is likely that some other constraints must be brought into play to account for the uniqueness of language in humans. Surprisingly, Arbib does not mention Miller's hypothesis that language is a product of sexual selection. Miller (2001) argues that the human brain must have been the kind of runaway process driven by sexual selection in a similar manner to Bower bird's nests and peacock's tails. Miller's arguments are extensively developed and show how Homo sapiens could have got the jump on very similar species with very similar brain architectures.

References

Ballard, D.H., Hayhoe, M.M., Pook, P.K., & Rao,R.P.N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*,1311-1328.

Miller, G. F. (2001). *The mating mind: how sexual choice shaped the evolution of human nature.* New York: Anchor Books.

Quine, W.V. (1960). *Word and object.* Cambridge, MA: The MIT Press.

Yu, C., & Ballard, D. H. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception, 1*, 57-80.

Yu, C., Ballard, D. H., & Aslin, R. N. (submitted). The role of embodied intention in early lexical acquisition. *Journal of Cognitive Science.*