# Multimodal Concepts for Social Robots

**Olivier Mangin** (Yale University; olivier.mangin@yale.edu)
DOI: 10.1145/3054837.3054844

Whether they try to mimic human cognition or give us a new glimpse of it as modelling tools, robots have become an essential component of cognitive sciences. My doctoral work embraces this duality of robotic research along two topics that are the study of language acquisition and learning by imitation.

How do concepts such as *red* or *grasp* emerge from our perceptual experience? When S. Harnad formulated the *symbol grounding problem* (Harnad, 1990), he shifted the traditional artificial intelligence's focus away from symbol manipulation problems to the definition, origin, and existence of these symbols. I hence study that question in line with two fundamental perspectives from developmental robotics: intelligence is the result of a developmental process and intelligence is embodied, i.e. it happens in the context of a sensorimotor interaction between the body and the world (Steels, 2008). How do we learn to recognize the words '*red*' and '*grasp*'? How do we relate the action of grasping performed by ourselves and someone else? Despite their apparent similarity to the first one, these two additional questions were originally studied in isolation, by distinct communities. One of the main contribution of my doctoral research is to bring these problems together in an interdisciplinary approach. To do so, I frame them into a common model: *the emergence of concepts from perception*, that covers a definition of motion primitives, the acquisition of words, as well as the discovery of objects in vision.

How can an animal or a robot acquire and represent skills, so that it can later reuse them and combine them into new more complex skills? My work focuses on the ability for robots to decompose demonstrations of complex motions or skills into a repertoire of primitive gestures, an approach that is inspired on the idea of *motion primitives*, rooted on biological evidences (Mussa-Ivaldi & Bizzi, 2000). It has promising applications for robots and in particular robots programmed by demonstration, an approach inspired from imitation learning in humans, to teach new skills to robots without needing expert programming. Indeed, robot programming by demonstration is currently limited to simple skills and is expected to greatly benefit from the ability to relate complex skills to a growing repertoire of primitive ones (Cangelosi, A. et al., 2010).

My doctoral research brings the following original contributions. First, I focused on the simultaneous combination of motion primitive instead of the more commonly studied sequential one. As a starting point, I built datasets of dance motions, that are short choreographies obtained by the simultaneous mixture of basic gestures. I additionally developed experiments and devised algorithms to learn gestures together with words (symbolic or from continuous speech) that describe the gestures (Mangin & Oudeyer, 2012, 2013). Indeed, social conventions play an important role in the perception of gestures which makes the motion decomposition problem alone ill-posed. Interestingly, in my thesis (Mangin, 2014), I frame the process of learning words, gestures, as well as visual objects, into a multimodal problem that treats all modalities similarly. The fact that one single algorithm enables both the learning of words and gestures establishes a strong symmetry between these problems. As another significant consequence, this experiment shows that, although these three individual problems are ambiguous, rephrasing them together as a multimodal learning problem can overcome their individual ambiguities. In particular, it allows the acoustic concepts of words to both shape and be shaped by the visual model of objects or gestures (Mangin, Filliat, Bosch, & Oudeyer, 2015), thus providing a model of the co-organization of language and concepts (Lupyan, Rakison, & McClelland, 2007).

Finally, these models are based on a simple compression heuristics and operate on unsegmented sentences (raw acoustic signal) in a multimodal sensory flow, without enforcing an *a priori* structure on sentences. Therefore this research enables a bottom-up approach to a

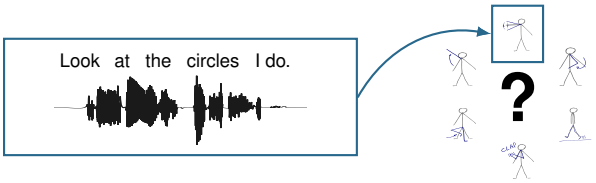


Figure 1: We evaluate the learner's behavior on a multimodal classification task: it hears a new utterance (raw acoustic signal) and choses the best matching gesture among demonstrations.

problem mostly studied from a top-down perspective. Moreover, I designed experiments to elicit *a posteriori* the acquisition of structure; these use behavioral tasks analogous to a child associating words to objects rather than requiring to look for some explicit representation in the robot's brain (see Figure 1).

This work presents an artificial learner that acquires word knowledge from cross-situational information. Future work includes experiments on realistic data to better compare such information in word learning to other cues such as mutual exclusivity or whole object assumption (Markman, 1990). Another direction is the learning of higher levels of structure, as grammar structures for language but also gestures. Finally incremental learning algorithms would enable to study the developmental path of the learning of such structure.

In summary, this research manages to bring together questions coming from several different domains and demonstrates that these questions can shed light on each other. In particular I posit an *algorithmic analogy* between the questions, that includes applications of well known algorithms (nonnegative matrix factorization) to new domains (such as motion decomposition), as well as the development of new algorithms (matrix factorization for inverse reinforcement learning). To conclude, I would like to stress the effort that has been achieved to make data, code[1], and results openly available for dissemination and for reproduction of the experiments.

## References

Cangelosi, A. et al. (2010). Integration of Action and Language Knowledge: A Roadmap for Developmental Robotics. *IEEE TAMD*, *2*(3), 167–195.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083.

Mangin, O. (2014). *The emergence of multimodal concepts: From perceptual motion primitives to grounded acoustic words* (PhD thesis). Université de Bordeaux.

Mangin, O., Filliat, D., Bosch, L., & Oudeyer, P.-Y. (2015). MCA-NMF: Multimodal concept acquisition with non-negative matrix factorization. *PLoS ONE*, *10*(10).

Mangin, O., & Oudeyer, P.-Y. (2012). Learning to recognize parallel combinations of human motion primitives with linguistic descriptions using non-negative matrix factorization. In *IROS*. Vilamoura, Portugal.

Mangin, O., & Oudeyer, P.-Y. (2013). Learning semantic components from subsymbolic multimodal perception. In *ICDL-EpiRob*.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57–77.

Mussa-Ivaldi, F. A., & Bizzi, E. (2000). Motor learning through the combination of primitives. *Philosophical trans., Royal Soc. of London.*, *355*(1404), 1755–69.

Steels, L. (2008). The symbol grounding problem has been solved. so what's next? In *Symbols and embodiment* (chap. 12). Oxford University Press.

[1] http://github.com/omangin/multimodal

**Olivier Mangin** is a developmental and social roboticist, now a postdoct in Prof. Scassellati's laboratory at Yale University. His interest lies in the apparition of structure in the sensorimotor interaction between an infant or robot and its environment. In particular he studies the development of linguistic skills, and the acquisition and understanding of behaviors during human-robot collaboration. O. Mangin completed his PhD at INRIA in France, under the supervision of P.-Y. Oudeyer.