Journal of Neurolinguistics 25 (2012) 500–514

Contents lists available at ScienceDirect

Journal of Neurolinguistics

journal homepage: www.elsevier.com/locate/jneuroling

The syntax of human actions and interactions

Gutemberg Guerra-Filho [a,*], Yiannis Aloimonos [b]

[a] Department of Computer Science and Engineering, University of Texas at Arlington, Nedderman Hall, Arlington, TX 76019, USA
[b] Department of Computer Science, University of Maryland at College Park, A.V. Williams Building, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 31 October 2008
Received in revised form 24 November 2009
Accepted 7 December 2009

Keywords:
Sensory-motor linguistics
Human Activity Language
Concept grounding

ABSTRACT

Human motion is a natural phenomenon that involves several different aspects in the representational level. Among these aspects, we find the discovery of motor primitives used to build complex motion; the representation of complex actions in terms of these primitives; the generalization of movement concerning different parameters such as target location, speed, and resistance force; the temporal concatenation of motion in a sequence of actions that considers co-articulation; and the parallelization of movement in space that allows the performance of different actions at the same time (*e.g.*, walk and wave). In order to model all important aspects of human motion, we seek a representation that considers these problems in a single framework. This way, we advocate that human motion may be represented as a language. Our Human Activity Language (HAL) consists of kinetology, morphology, and syntax. Kinetology, the phonology of human movement, involves the learning of motor primitives through segmentation and symbolization. Morphology concerns the representation of action words in terms of kinetemes and the discovery of a set of essential coordinated actuators for each action. Syntax is related to the construction of motion sentences using action words in sequence or in parallel. In this paper, we extend HAL syntax to consider human interactions between two subjects. We captured a praxicon, lexicon of human movement, with a number of human interactions such as shake hands, shove, and pass on. We empirically show that human interactions have a particular syntax based on the syntax of individual actions.

* Corresponding author.
E-mail address: guerra@cse.uta.edu (G. Guerra-Filho).

0911-6044/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jneuroling.2009.12.006

1. Introduction

In the sensory pathway, the cognitive understanding of human activities involves the analysis (parsing) of observed action sequences according to an organized *praxicon*, a structured lexicon of human actions, previously learned and stored. In the motor pathway, the cognitive process concerns the synthesis (generation) of action sequences based on this praxicon. A sensory-motor praxicon is organized according to some knowledge representation. We advocate a linguistic representation to support artificial cognitive systems on the purpose of motion synthesis and analysis.

Inspiration for a linguistic approach to human activity representation comes from converging evidence in several fields of science such as cognitive science, neuroscience, neurophysiology, and psychophysics. Observations and dissections of the brains of people with brain injuries and diseases have shown that areas of anterior and parietal cortex in the left hemisphere of the cerebrum provide control for both vocal and manual activity (Greenfield, 1991; Kimura, 1981; Poizner, Klima, & Bellugi, 1987). These activities include the hierarchical organization of manual combination of objects, signing, and speech. In addition to that, Broca's region in the human brain has functions related to language tasks. Broca's region also contributes to functions ranging from action perception to generation (Nishitani, Schurmann, Amunts, & Hari, 2005). The evidence of such a region in the brain with language and action functions is another inspiration for a linguistic approach to the representation of human activities.

A linguistic framework for a common sensory and motor representation is a reasonable approach since there is evidence that spoken language is semantically grounded also in action (Glenberg & Kaschak, 2002; Nishitani et al., 2005). At a higher level, the existence of mirror neurons (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996) in humans suggests that the same representation for motor information related to body movement is also used in the brain for perceptual tasks. The theory of motor tapes (Hoyle, 1983) is another theory where explicit representations of a movement trajectory are stored in memory. When an agent needs information on how to perform an action, it finds the appropriate template in memory and executes it.

Similarly to spoken language, movement patterns are composed of primitive elements in combination and sequences, but they may not be structured exactly like spoken language since dimensions are qualitatively different (Armstrong, Stokoe, & Wilcox, 1995). Humans make finely controlled movements that produce invisible (or barely visible) but audible gestures in the vocal tract (*e.g.*, throat and mouth). The information about these movements is broadcast to the environment for the purpose of communication. This way, speech can be characterized as interleaved patterns of movements, coordinated across several different articulators in the vocal tract (Studdert-Kennedy, 1987). The description of acoustic and visual gestures uses the same vocabulary of neuromuscular activity as a generalization of the vocal tract grammars at phonological level. Visible gesture words or sentences could have provided the behavioral building blocks associated with neuronal group structures for constructing syntax incrementally and neurologically (Edelman, 1992).

Language and visible movement use a similar cognitive substrate based on the embodiment of grammatical processing. For example, during walking acquisition, the human infant follows a developmental sequence that is similar to the sequence followed in language acquisition. The similarity of this developmental sequence is due to more basic underlying bio-behavioral forces and constraints. Stages in motor development reflect neuromuscular maturation. The fundamental stages of sign language and spoken language acquisition are the same (Volterra & Erting, 1990). Infants go through a babbling stage, in which they manipulate the sublexical elements (Petitto & Marentette, 1991). Language develops through social interaction since a word meaning is learned when heard or seen used by someone else in a context that made the relation between word and meaning reasonably unambiguous. Once language is acquired at a sufficient level, the meaning of unfamiliar words is determined by linguistic inference from its context.

Motivated by these evidences above, we propose a linguistic framework for the modeling and learning of human activity representations. Our ultimate goal is to discover a sensory-motor language, denoted as *Human Activity Language*, which represents the sequential and parallel aspects of human movement with perceptual and generational properties. Our approach finds a linguistic structure for human movement with analogs of phonology, morphology, and syntax.

A language for human activity has impacts in many areas. A symbolic representation for human activity materializes the concept of motor programs and enables the identification of common motor subprograms used in different activities. This way, the discovery of such language allows exploring how a motor activity vocabulary is organized in terms of its subprograms. More specifically, the usefulness of the linguistic methodology for human motion representation is as a tool to analyze and synthesize human movement. Our linguistic framework may be used to study motion disorders and motion-related diseases. For example, a linguistic representation may be applied to the evaluation of Parkinson's disease patients with regards to their response to a specific treatment. Movement analysis with such a tool could also lead to the early diagnose of autism, schizophrenia, and other diseases that show signs through the motor system. More usual applications are in the evaluation of the performance and injury recovery of athletes. We also should mention the tremendous potential for Computer Science in areas such as Computer Vision (surveillance with action recognition), Computer Graphics (motion-based animation), and Robotics (learning motor skills through imitation).

Our Human Activity Language (HAL) consists of kinetology, morphology, and syntax. *Kinetology* (Guerra-Filho & Aloimonos, 2006), the phonology of human movement, finds basic primitives for human motion (segmentation) and associates them with symbols (symbolization). This way, kinetology provides a symbolic representation for human movement that allows synthesis, analysis, and symbolic manipulation. Kinetology also has applications to compression, decompression, and indexing of motion data.

Morphology is responsible for the representational construction of a vocabulary of actions to aid an artificial cognitive system with tasks related to perception and actuation (Guerra-Filho, 2009). In order to learn action morphemes and their structure, we presented a grammatical inference methodology and introduced a parallel learning algorithm to induce a formal grammar system representing a single action. This process is performed for every action in a vocabulary of human actions and the praxicon is built this way.

In this paper, we discuss the syntax of human actions with a special interest on interactions between subjects. The syntax of human activities involves the construction of sentences using action morphemes obtained from the morphological system. Sets of morphemes represent simultaneous actions (parallel syntax) and a sequence of morphemes is related to the concatenation of activities (sequential syntax). Parallel syntax concerns the combination of motions of different body parts such as walk and wave at the same time. Sequential syntax is proposed as a modeling tool for the transitioning between different movements. A transition is a segment of motion that seamlessly attaches two motions to form a single longer sequence of motion. The sequential syntax allows the production of sentences that represent motion sequences such as stand up, then walk, and then jump.

The paper is organized as follows: automatic language learning is based on formal representations of language. In Section 2, we provide the necessary background on our linguistic framework. Since the syntax in our Human Activity Language builds on kinetology and morphology, we present a condensed review of these subjects. The reader is referred to our previous work for further details (Guerra-Filho, 2009; Guerra-Filho & Aloimonos, 2006). An extension of our linguistic framework to multi-subject interactions is also presented in this section. Section 3 presents the syntax of human activities including nuclear, parallel, and sequential syntaxes. A brief discussion and our conclusions are presented in Section 4.

2. Human Activity Language

Knowledge of actions is essential to human survival. Hence, infants acquire actions by observing and imitating the actions performed by others. Once basic actions are acquired, they learn to combine and concatenate simple actions to form more complex actions. This process can is similar to speech, where we combine phonemes into words, and words into sentences. Humans can recognize as well as generate both actions and speech. In fact, the binding between the cognitive and generative aspects of actions is revealed at the neural level in the monkey brain by the presence of mirror neuron networks, *i.e.*, neuron assemblies which are activated when the individual observes a goal-oriented action (like grasping) and also when the individual performs the same action. All these observations lead us to

a simple hypothesis: actions are effectively characterized by a language. This is a language with its own building blocks (phonemes), its own words (morphemes), and its own syntax.

The realm of human actions may be represented in at least three domains: visual, motor, and linguistic. The visual domain covers the form of human actions when visually observed. The motor domain covers the underlying control sequences that lead to observed movements. The linguistic domain covers symbolic descriptions of natural (*i.e.*, ecologically valid) actions. We take the hierarchical structure of natural language (*e.g.*, phonology, morphology, syntax) as a framework for structuring not only the linguistic system that describes actions, but also the motor system. We defined and computationally modeled motor control structures that are analogous to basic linguistic counterparts: phonemes (the alphabet), morphemes (the vocabulary), and syntax (the rules of combination of entries in the vocabulary) using data-driven techniques grounded in actual human movement data. Since actions have a visual, motor and a natural language, converting from one space to another becomes a language translation problem.

2.1. Human Activity Repository

What does it really mean to learn a language? According to modern linguistics, this amounts to learning the phonology, morphology, and syntax (and semantics/pragmatics) of the language. The input for our phonological system is a corpus of real human motion. Using motion capture equipment, we acquired motion data for hundreds of actions associated with English verbs related to observable voluntary meaningful movement.

We used an optical motion capture system with 16 cameras at 120 frames per second. The cameras are evenly placed in a circular configuration and four different height levels looking at the center of the capture volume. Our capture volume is approximately 8 ft (length) $\times$ 6 ft (width) $\times$ 8 ft (height), which allows for 10 ft locomotion along the diagonals. A number of 39 spherical retro-reflective markers are placed on the body skin of the subjects at joint articulations and other salient anatomic places. The motion capture system finds the location of these markers in Cartesian space (x, y, z coordinates) and all degrees of freedom of an articulated body model (global location, orientation, and joint angles) are retrieved.

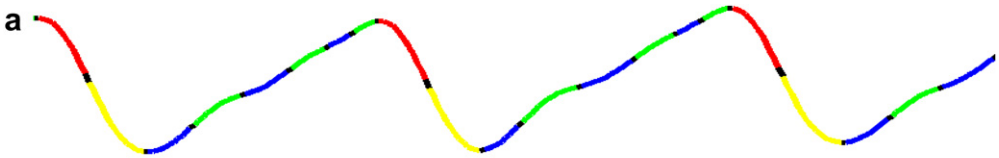
We collected a Human Activity Repository (HAR) with about 500 actions so far. Our repository consists of 350 actions involving a single individual (subject *A*) and 150 actions involving two subjects (subjects *A* and *B*) interacting at the same time. Each action is performed by the subjects repeatedly and consistently for at least 10 consecutive times in a single motion sequence. Consistency in this case means that the subject gives the best effort to repeat each action in the very same manner and speed. Each motion sequence corresponds to a file in our repository.

The articulated human body is represented by a hierarchy of rigid parts connected by articulated joints. The root of this skeleton is associated with 6° of freedom (DOF) to describe the global position and orientation in the world coordinate system. Each articulated joint corresponds to up to three DOFs associated with rotational angles. Each DOF, denoted as *actuator*, is modeled as a time-varying function $M_i(t)$, where i denotes the degree of freedom and t is the time frame of the motion sequence.

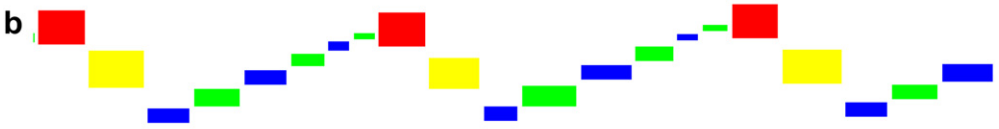
2.2. Kinetology

Since our object of study is motion, not sound, phonology takes the name of kinetology in our approach and the phonemes become *kinetemes*. In kinetology, our goal is to identify the motor primitives (segmentation) and to associate them with symbols (symbolization). This way, kinetology provides a grounded symbolic representation for human movement. While *motion synthesis* is performed by translating the symbols into motion signal, *motion analysis* uses this symbolic representation to transform the original signal into a string of symbols.

In order to find motion primitives, each joint angle function is divided into segments. The *segmentation* process starts by assigning a state to each instant (see Fig. 1a,b). Adjacent instants assigned to the same state belong to the same segment. The state is chosen from a set that includes all possible sign combinations of angular derivatives (*i.e.,* velocity, acceleration, and jerk). For example, considering angular velocity (M') and angular acceleration (M''), there are four possible states: $\{M'_i(t) \geq$

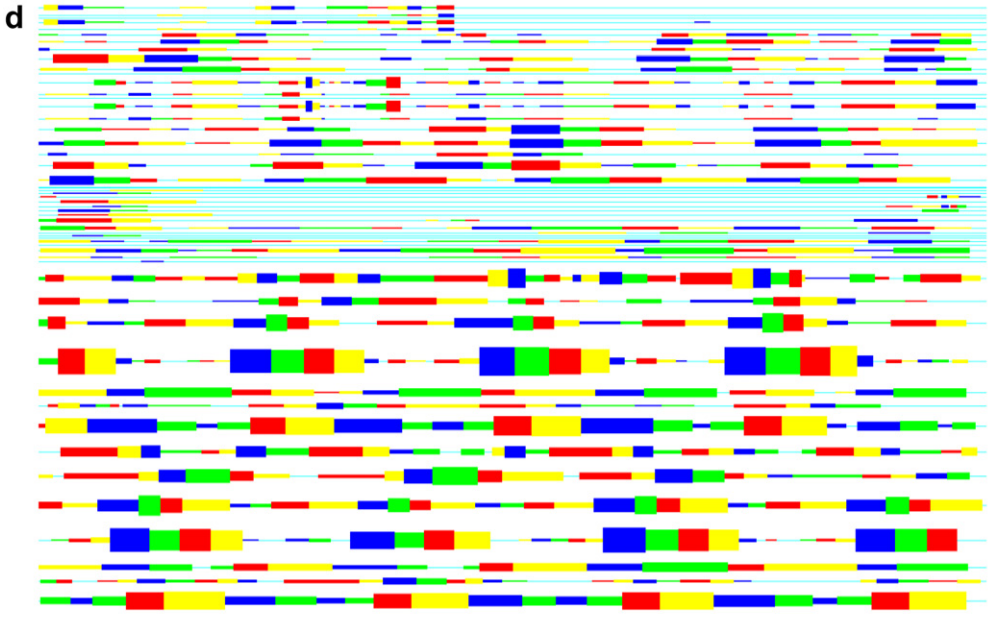


Geometric representation: a joint angle function with assigned states corresponding to positive velocity and positive acceleration (blue); positive velocity and negative acceleration (green); negative velocity and positive acceleration (yellow); and negative velocity and negative acceleration (red).



Segmentation: the sequence of kinetemes for the joint angle function above.

Symbolization: symbols are associated with each kineteme of the segmented sequence.



Actiongram: a set of symbolic strings for each actuator in the body model.

Fig. 1. A four-state kinetological system.

0 and $M''_i(t) \geq 0$; $M'_i(t) \geq 0$ and $M''_i(t) < 0$; $M'_i(t) < 0$ and $M''_i(t) \geq 0$; $M'_i(t) < 0$ and $M''_i(t) < 0\}$. These states are depicted with blue, green, yellow, and red colors respectively in the figures below. In the case where we consider only the angular velocity, there are only two possible states $\{M'_i(t) \geq 0; M'_i(t) < 0\}$ shown with blue and red colors respectively.

Once the segments are identified, we keep three attribute values for each segment: the state, the angular displacement (*i.e.*, the absolute difference between initial angle and final angle), and the time

period length. Each segment is graphically displayed as a filled rectangle, where the color represents its state, the vertical width corresponds to angular displacement, and the horizontal length denotes the time period length. Given this compact representation, the attributes are used in the reconstruction of an approximation for the original motion signal and in the symbolization process.

The *symbolization* process consists in associating each segment with a symbol such that segments corresponding to different performances of the same motion are associated with the same symbol (see Fig. 1c). In the context of finding patterns in time-series data, there is a vast literature. A simple method to solve this problem is hierarchical clustering. This approach is based on a distance metric that measures the similarity between different segments.

Given the segmentation for a motion sequence, the symbolization output is a set of strings which defines a data structure denoted as *actiongram* (see Fig. 1d). An actiongram A has n strings $A_1, \ldots, A_n$. Each string A_i corresponds to an actuator of the human body model and contains a (possibly different) number of n_i symbols. Each symbol $A_i(j)$ is associated with a segment.

Our algorithms will work with thousands of actiongram representations, such as the one in Fig. 1d, with the goal of finding further structure in these kineteme sequences (*i.e.*, yellow, red, blue, and green rectangles of various sizes). There are two kinds of structure: (a) along each actuator and (b) among different actuators (coordination or synergies). These two structures are the ones we learn from real data in our morphological process.

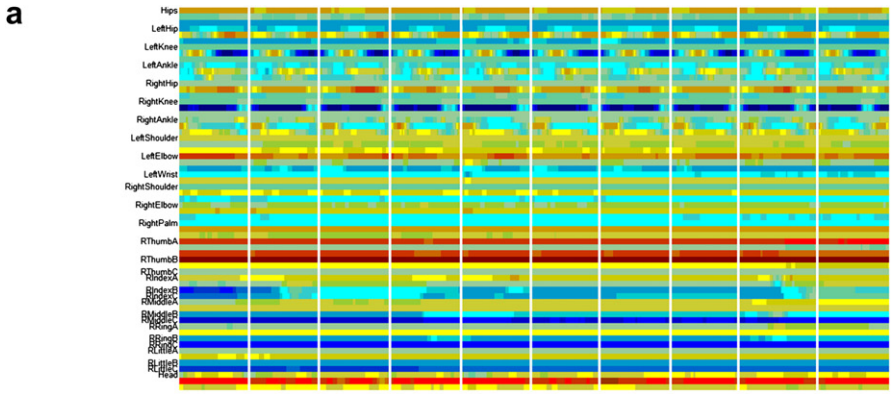
2.3. Morphology

The morphology of a human action is related to the essential parts of the movement and its structure. A single action morpheme represents the least amount of movement with a purposeful goal, *i.e.*, meaning. In this sense, we define a human action *morpheme* as the set of essential actuators intrinsically involved in the action and the corresponding motion patterns (in terms of kinetemes). The morphemes are the essential parts of human actions. Since the derived motion patterns are sequences of kinetemes, the inference of morphemes is called *morpho-kinetology*. This part of morphology aims to select a subset of the motion which projects the whole action only into the essential actuators and their motion patterns (see Fig. 2).

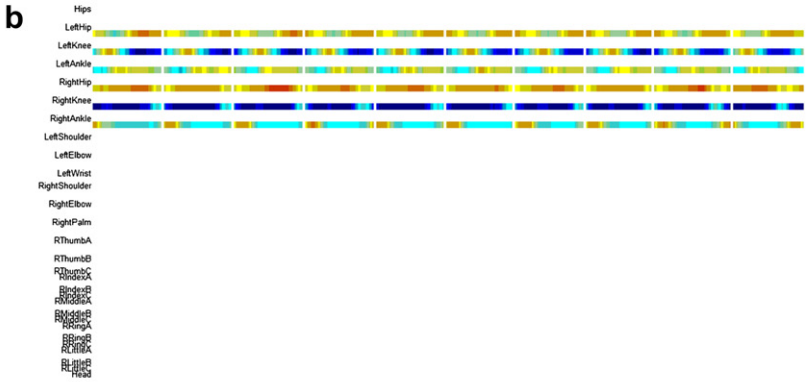
The *essential actuators* are the ones actually responsible for the achievement of the intended result of an action. They are strongly constrained and, consequently, only these "meaningful" actuators will have consistent motion patterns in different performances of the same action. Therefore, the inference of the morpheme of a human action requires an actiongram that contains several executions of the same action. Given such an actiongram A as input, we aim to automatically learn the morpheme of the corresponding action. Formally, the morpheme consists of a set I representing the essential actuators for the action and, for each $i \in I$, a substring p_i of A_i corresponding to the motion pattern that the actuator i performs during the action. Since our input is a set of strings, we pose this problem as a grammatical inference.

Grammatical inference concerns the induction of the grammar of a language from a set of labeled sentences. The grammar inference consists in learning a set of rules for generating the valid strings that belong to the language. The target grammar usually belongs to the Chomsky hierarchy of formal grammars. There exist several methods for learning regular grammars, context-free grammars (CFGs), and stochastic variations.

We denote *sequential learning* as the technique able to infer the structure of a single sequence of symbols A_i. This structure corresponds to a forest of binary trees (see Fig. 3), where each node in a tree is associated with a context-free grammar rule in a normal form. Initially, the sequential learning algorithm computes the number of occurrences for each different digram in the string A_i. A *digram* is a pair of adjacent symbols. A new grammar rule $N_c \rightarrow \alpha\beta$ is created for the digram $\alpha\beta$ with the current maximum frequency. The algorithm replaces each occurrence of $\alpha\beta$ in the string A_i with the created non-terminal N_c. The whole procedure is repeated until digrams occur more than once. As an example, the set of rules inferred for the Context-Free Grammar (CFG) displayed in Fig. 3 is: $\{N_1 \rightarrow AB, N_2 \rightarrow CD, N_3 \rightarrow EF, N_4 \rightarrow BN_1, N_5 \rightarrow N_2N_3, N_6 \rightarrow N_5G, N_7 \rightarrow N_6N_4\}$. A sequential learning algorithm keeps merging adjacent root nodes into single rules and, consequently, over-generalization happens when "unrelated" rules are generalized.



The whole motion for all actuators. It represents human movement data for ten performances of the same action. The labels are names of joints (*e.g.*, left hip, left knee). For each joint, we have at most three Euler angles in the measurements coded with color.



The whole motion described only by the essential actuators. The other actuators are irrelevant.

Only the motion patterns for the essential actuators. This is a single morpheme.

Fig. 2. The morpho-kinetology process selects a subset of the motion which projects the whole action only into the essential actuators and their motion patterns.

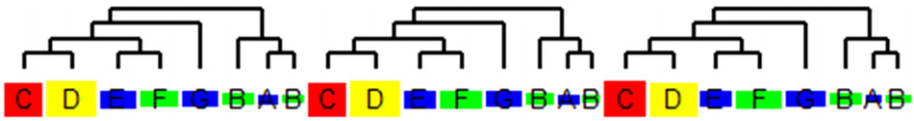


Fig. 3. Sequential learning on a single string of an actiongram. Here we show how we can start from the sequence of kinetemes in a joint and learn a grammar (tree) that generates the string.

We proposed *parallel learning* to concurrently infer a *grammar system*, a set G of context-free grammars related by synchronized rules (see Fig. 4), as the structure of all strings $A_1, \ldots, A_n$ in the actiongram A. Each string A_i in an actiongram corresponds to the language which will be inferred for a component G_i modeling an actuator i. Our *parallel learning* algorithm executes the sequential learning within each string A_i independently. In parallel learning, nodes are merged only if the new rule is synchronized with other rules in different CFG components of a grammar system. This way, over-generalization is avoided since synchronization guarantees a relationship between the merged rules.

Once morphemes are inferred for each action in a praxicon, we may learn further structure for these morphemes. This structure arises from the ordering, intersection, and repeated occurrences of kinetemes in motion patterns for the same actuator but in different actions. We refer to this additional structure as *morpho-syntax*.

Our method to infer morpho-syntax considers a single actuator i at a time. We denote p_i^a as the motion pattern for actuator i and action a, such that $i \in I^a$, where I^a is the set of essential actuators for action a. Basically, all motion patterns p_i^a for actuator i in different actions are described as sequences of kinetemes. These sequences altogether can be generated by a single context-free grammar that represents a more compact and efficient structure: a morphological grammar. The syntax of human activities is based on this morphological grammar that represents a structured praxicon.

2.4. Multi-subject interactions

Human interactions consist of actions that involve more than one subject to accomplish the same goal, collaborative goals, or antagonistic goals. A multi-subject interaction is an activity where a set $S = \{s_1, s_2, \ldots, s_k\}$ of k subjects with different roles interact towards related goals. A multi-subject

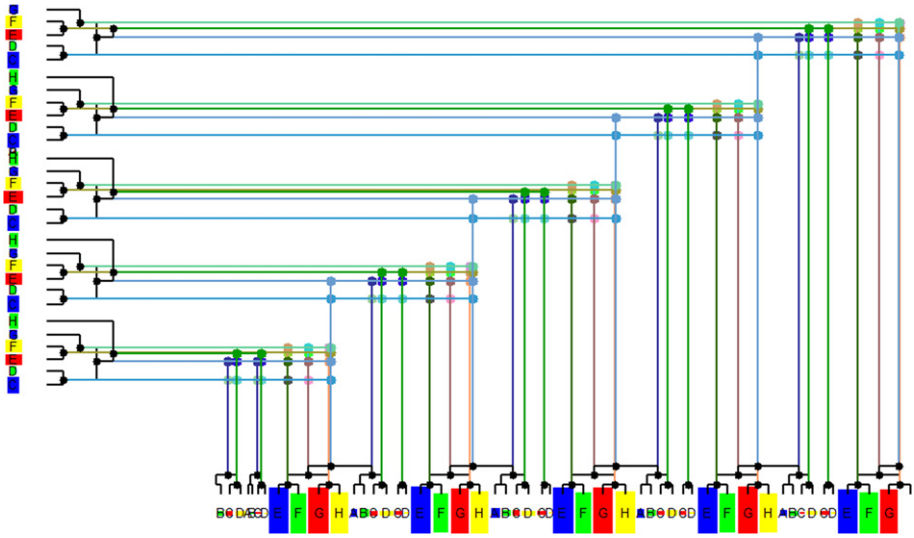


Fig. 4. Two CFGs, binary trees corresponding to hip (vertical) and knee (horizontal) flexion/extension, of a PSGS for the action walk related by synchronized rules (colored lines) which represent the synergies.

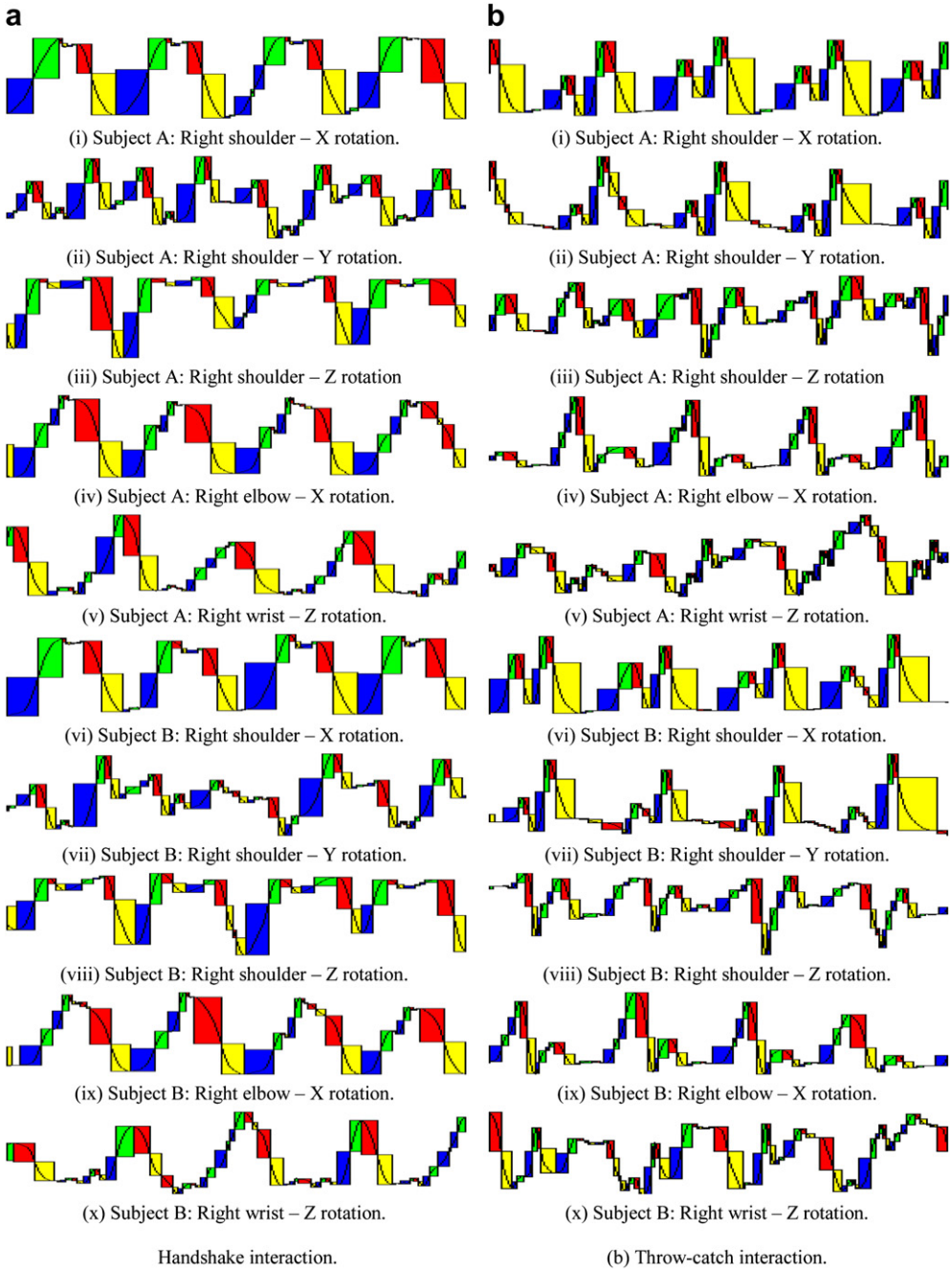


Fig. 5. Two interactions between two subjects represented by five essential actuators for each subject.

interaction is described by the structure of the motion of every pair (s_u, s_v) of subjects s_u and s_v in the set S of interacting subjects. This way, the morpheme of a k-subject interaction is represented by $k(k-1)/2$ morphemes of two subject interactions. Without loss of generality, we consider interactions that involve only two subjects: subject A and subject B. Fig. 5 shows two examples of interactions between

two subjects, namely the handshake action and the throw–catch action, where five essential actuators for each subject are displayed.

To consider two subject interactions, the set of actuators is augmented to contain the degrees of freedom of both subjects *A* and *B*. The morpheme of a two-subject interaction is obtained by parallel learning in the same way as a single subject action morpheme is found. The difference is that the morphological process in this case finds a set of essential actuators that contains actuators belonging to both subjects. Note that the interaction has some level of coordination between the two subjects and, consequently, the actuators of the different subjects are also coordinated. For example, the subjects may interact synchronously, as in the case of a handshake, or move according to an action-reaction pattern where a subject responds to the others action, as in the case of a throw–catch interaction. The coordination is inferred by parallel learning which detect synchronous rules between actuators of the same subject and also between actuators of the different subjects.

3. Syntax

The syntax of human activities involves the construction of sentences using action morphemes. A sentence consists of a group of entities. In this sense, a sentence may range from a single action morpheme to a sequence of sets of morphemes. The sets of morphemes represent simultaneous actions and a sequence of movements relates to the causal concatenation of activities. This way, our intention is to identify which entities constitute a single morpheme sentence (nuclear syntax) and to study the mechanisms of composing sets of morphemes (parallel syntax) and of connecting these sets into sequences (sequential syntax).

3.1. Nuclear syntax

A single action morpheme sentence is composed of entities that are implicit in any motion. These entities are a central part of an action that we refer as *nuclear-syntax*. For didactical purposes, we identify these entities as analogs to lexical categories: nouns, adjectives, verbs, and adverbs. An action is represented by a word that has the structure of a sentence: the agent or subject is a set of active body parts (noun), and the action or predicate is the motion of those parts (verb). In many such words, the action is transitive and involves an object or another patient body part.

3.1.1. Nouns and adjectives

In a sentence, a noun represents the subjects performing an activity or objects receiving an activity. A *noun* in a single action sentence corresponds to the essential body parts active during the execution of a human activity and to the possible objects involved passively in the action (including patient body parts). The body parts are equivalent to actuators of the articulated body model. Therefore, a noun (active body parts) is retrieved from the set of essential actuators in the action morpheme. This set may be represented as a binary string with the same size of the set of all actuators. Each element of this string encodes the inclusion of a particular joint actuator in this set. Given the morphology of each action in our motion corpus, we may find a matrix where each column is a binary string encoding the noun for a different action (see Fig. 6). This way, the rows of this matrix correspond to actuators. The *noun matrix* is a low-level structure containing the vocabulary of nouns for a praxicon.

The initial posture of an action is analogous to an *adjective* which further describes (modifies) the active body parts (nouns) in the sentence. The initial pose of an action is retrieved from a morpheme as the initial joint angle of the first kineteme in the motion pattern of each essential actuator. This way, adjectives represent usual initial postures such as sitting, standing, and lying.

3.1.2. Verbs and adverbs

A motion *verb* represents the changes each active actuator experiences during the action execution. The human activity verbs are obtained from the motion patterns in the action morphemes.

An *adverb* models the variation in the execution of each motion segment in a verb. The adverb modifies the verb with the purpose of generalizing the motion. For example, an instance of a "shake hands" interaction corresponds to a morpheme that represents the motion required to reach, grab, and

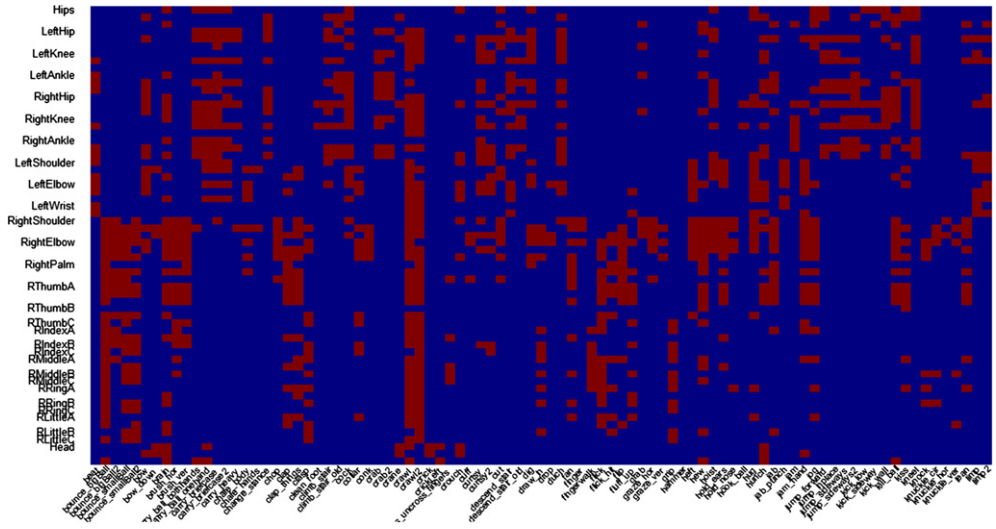


Fig. 6. Noun matrix for a praxicon.

shake someone's hand located at a specific target location. To generalize this action to any location, the motion of a segment is represented in a parametric space with a reduced dimensionality. Each dimension in this reduced space represents a parameter such as target location, speed, and resistance force that models the variability of an action or interaction. We use an intuitive set of parameters that consist of origin, destination, speed, and a resistance force. For example, the "shake hands" action is specified by the start position of the hand (possibly at a resting location), the end position of the hand (where the other hand is located at), how fast the movement is performed, and any resistance force involved in the interaction. The adverbial modeling process consists in the mapping from parametric space to motion. Consequently, this process provides an immediate mapping between extrinsic coordinates (Cartesian 3D locations) and intrinsic coordinates (joint angles). The modeling process involves the interpolation of sample motions in the parametric space.

3.2. Spatio-temporal syntax

The organization of human movement is also simultaneous and sequential. This way, the nuclear syntax expands to parallel and sequential syntax. The parallel syntax concerns activities performed simultaneously represented by parallel sentences $S_{t,j}$ and $S_{t,j+1}$. This syntax constrains the respective nouns of the parallel sentences to be different: $N_{t,j} \neq N_{t,j+1}$. This constraint states that simultaneous movement must be performed by different body parts. For example, a person may walk and wave concurrently. However, one cannot whistle and chew gum at the same time!

The temporal sequential combination of action sentences ($S_{t,j}\, S_{t+1,j}$) must obey the cause and effect rule. The HAL noun phrase must experience the verb cause and the joint configuration effect must lead to a posture corresponding to the noun phrase of the next sentence. Considering noun phrases as points and verb phrases as vectors in the same space, the cause and effect rule becomes $NP_{t,j} + VP_{t,j} = NP_{t+1,j}$. The cause and effect rule is physically consistent and embeds the ordering concept of syntax.

3.2.1. Parallel syntax

Parallel syntax addresses the possible ways to combine different action morphemes into a set of morphemes that could be performed simultaneously. Basically, the main constraint imposed by parallel syntax involves the essential actuators. To merge two action morphemes for actions a_1 and a_2 into a parallel set of morphemes, their sets of essential actuators I^{a_1} and I^{a_2} need to have an empty

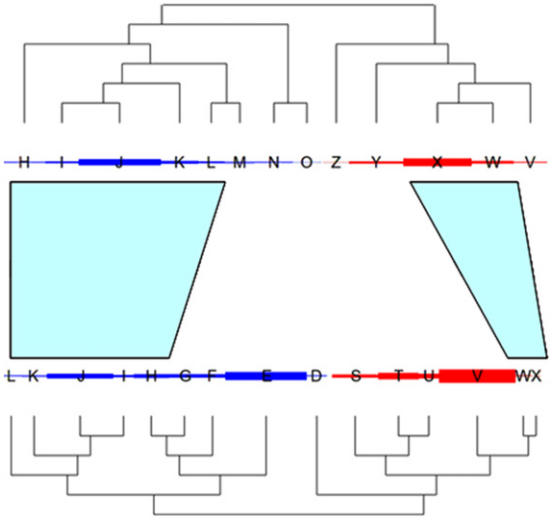


Fig. 7. Possible transitions between two morphemes.

intersection. In other words, the two action morphemes cannot share any essential actuator. This rule may be implemented as a Boolean constraint matrix C. For each pair of actions a_1 and a_2 in a praxicon, if $I^{a_1} \cap I^{a_2} = \varnothing$, the matrix entry $C(a_1, a_2)$ is true; otherwise, the matrix entry is false. The constraint matrix explicitly stores which pairs of morphemes could be merged as simultaneous activities. More sophisticated inferences could also be performed using this structure. For example, transforming this matrix into a graph, cliques correspond to groups of action morphemes that may be executed at the same time.

3.2.2. Sequential syntax

In speech, the temporal organization is a pre-syntax since this neural preplanning of motor action is what syntax uses to execute an utterance. Actions of the physical body provide a metaphor for the hierarchical structure of language. The precise muscle timing (pre-syntax) makes it possible to produce countless actions that differ in great or small ways. The lexical units are arranged into sequences to form sentences. A sentence is a sequence of actions that achieve some purpose.

The cause and effect rule is physically consistent and embeds the ordering concept of syntax. The body pose must experience the motion cause and the effect leads to a posture in the next sentence. Sequential syntax concerns the concatenation of actions or, more formally, the connection of sets of action morphemes (from parallel syntax) to form sequences of movement.

Consider a single actuator i, if i belongs to the sets I^{a_1} and I^{a_2} of essential actuators of two action morphemes a_1 and a_2, respectively, the sequential concatenation of these two morphemes is only feasible if there is a transition from one motion pattern p_{a_1} to the other p_{a_2}. Such a transition may be obtained from the morphological grammar G_i of actuator i. Any kinetemes shared by both motion patterns p_{a_1} and p_{a_2} give rise to a possible transition. Consequently, the two morphemes a_1 and a_2 have a *feasible concatenation* with respect to actuator i. This way, two sets of action morphemes may be sequentially connected only if they have a feasible concatenation with respect to all actuators contained in the intersection of their sets of essential actuators. Fig. 7 displays the motion patterns of two action morphemes and their respective morphological grammar entries. The two patterns share kinetemes and, consequently, a transition exists between the two morphemes. For example, a subject may transition from a walk to a run action. Fig. 7 shows the CFG grammar associated with the right knee flexion-extension of these two actions. For this specific joint angle, transitions are feasible using the sequences of kinetemes XWV and HIJKL. Therefore, a transition from walk to run is feasible with regards the right knee flexion-extension.

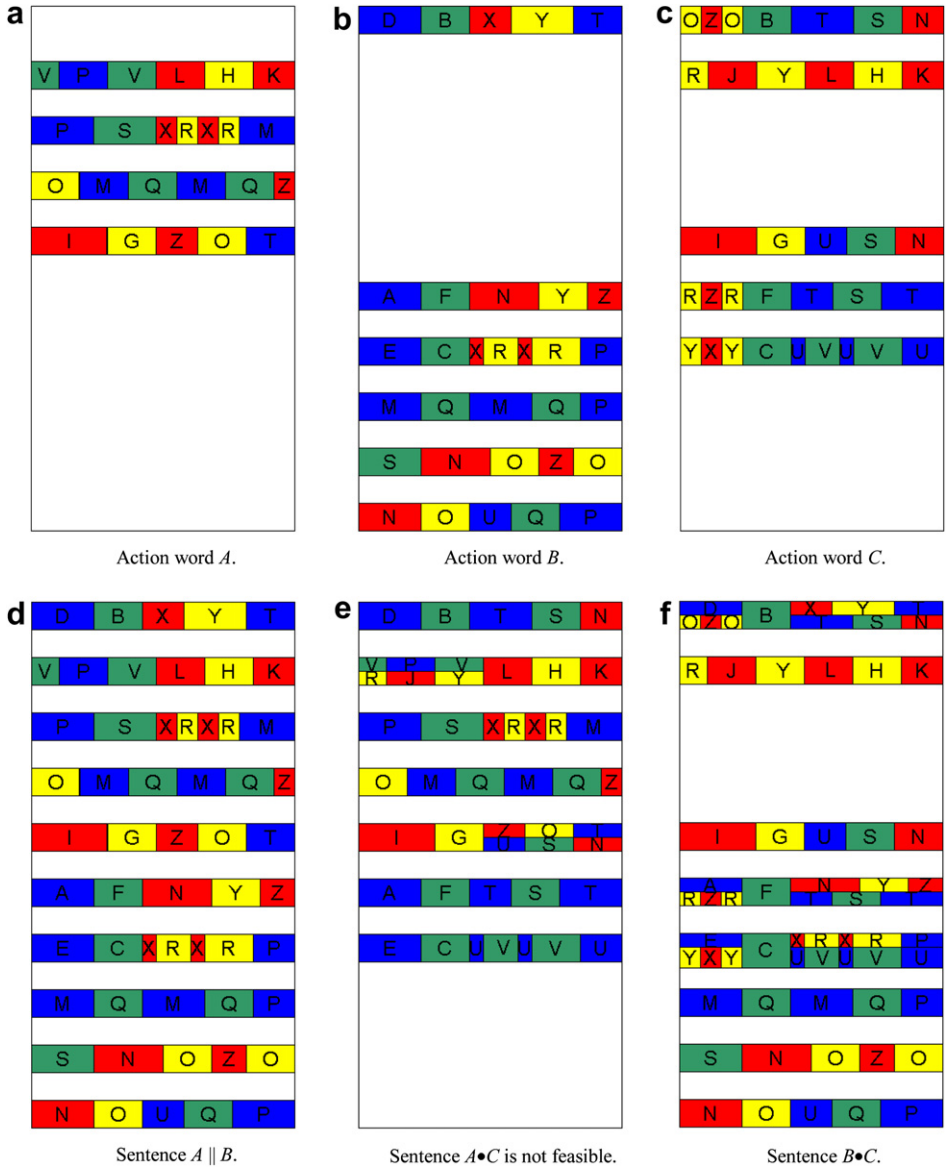


Fig. 8. Sentence formation process.

The lexical units are arranged into sequences to form sentences. A sentence is a sequence of actions that achieve some purpose. In written language, sentences are delimited by punctuation. Analogously, the action language delimits sentences using motionless actions. In general, a conjunctive action is performed between two actions, where a conjunctive action is any preparatory movement that leads to an initial position required by the next sentence.

In Fig. 8, we illustrate the sentence formation process using three action words *A*, *B*, and *C* (see Fig. 8a–c). Each horizontal row in an action word corresponds to a particular actuator. The motion pattern for each essential actuator is shown as a sequence of kinetemes depicted as colored rectangles with symbols. Since *A* and *B* have disjoint sets of essential actuators, we can form the simultaneous

sentence $A\|B$ (see Fig. 8d) using parallel syntax. For example, the actions walk and wave form a simultaneous sentence for a single subject. Another example is the sentence obtained with the interactions hold-a-box together and walk. They represent a simultaneous sentence for two subjects collaborating towards the objective of carrying a box together. On the other hand, action words A and C share two essential actuators and, consequently, only a sequential composition applies. Although there are transitions for both actuators, they are not concurrent (transition ⟨LHK⟩ occurs in a time period different from transition ⟨IG⟩) and the sequential sentence $A \cdot C$ is not feasible (see Fig. 8e). However, since transitions ⟨B⟩, ⟨F⟩, and ⟨C⟩ are concurrent, the sequential sentence $B \cdot C$ is feasible (see Fig. 8f). For a sequential composition, we find examples in transitions between different ways of locomotion (such as from walk to run as mentioned above and from run to a jump). These compositions are possible because there are feasible transitions for all common essential actuators of both actions. In terms of interactions, two subjects could concatenate the activities could lift-a-box, carry-a-box, and place-a-box into a single sequential sentence.

4. Conclusions

Perhaps, a more general cognitive capacity allows us to decode highly coded signals. A reasonable hypothesis would be that there is little difference between the visual and the speech realms in this regard. The visual stimuli available to the brain do not offer a stable code of information. The brain extracts the constant invariant features of objects from the perpetually changing flood of information it receives from them. Further, what is being perceived and apprehended is the message itself, not static target end states of the articulators.

We simply demonstrated that there exists a language of human activity by empirically constructing one such language out of large amounts of data. Our kinetology was among the simplest possible, yet rich enough to provide an interesting structure. It should be clear that there is a trade-off between the complexity of the kinetology and the complexity of the grammar. Very simple kinetemes give rise to complex grammars, while more structured kinetemes produce simpler grammars. A recent effort is to develop a spectral kinetology, where the kinetemes are basic functions (wavelets) linked with a number of parameters for each joint. The idea is that a single wavelet in conjunction with the provided parameters will produce the whole function (movement) of a synergy of joints. This approach will give rise to simpler grammars.

From a methodological viewpoint, this framework introduced a new way of achieving an artificial cognitive system through the study of human action, or to be more precise, through the study of the sensory-motor system. We believe this study represented initial steps of one approach towards conceptual grounding. The closure of this semantic gap will lead to the foundation of concepts into a non-arbitrary meaningful symbolic representation based on sensory-motor intelligence. This representation will serve to the interests of reasoning in higher-level tasks and open the way to more effective techniques with powerful applications.

Humans have been studying the spoken and written languages for thousands of years. It is not clear how long it will take to map out the murky depths of a Human Activity Language. We hope that HAL is a step in the right direction.

References

Armstrong, D., Stokoe, W., & Wilcox, S. (1995). *Gesture and the nature of language*. New York: Cambridge University Press.
Edelman, G. (1992). *Bright air, brilliant fire: On the matter of mind*. New York: Basic Books.
Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*(2), 593–609.
Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9*(3), 558–565.
Greenfield, P. (1991). Language, tools and brain: the ontogeny and phylogeny of hierarchically organized sequential behaviour. *Behavioral and Brain Sciences, 14*, 531–595.
Guerra-Filho, G. (2009). The morphology of human actions: finding essential actuators, motion patterns, and their coordination. *International Journal of Humanoid Robotics, 6*(3), 537–560.
Guerra-Filho, G., & Aloimonos, Y. (2006). Understanding visuo-motor primitives for motion synthesis and analysis. *Computer Animation and Virtual Worlds, 17*(3–4), 207–217.
Hoyle, G. (1983). *Muscles and their neural control*. New York: John Wiley.
Kimura, D. (1981). Neural mechanisms in manual signing. *Sign Language Studies, 33*, 291–312.
Nishitani, N., Schurmann, M., Amunts, K., & Hari, R. (2005). Broca's region: from action to language. *Physiology, 20*, 60–69.

Petitto, L., & Marentette, P. (1991). Babbling in the manual mode: evidence for the ontogeny of language. *Science, 251*, 1493–1496.
Poizner, H., Klima, E., & Bellugi, U. (1987). *What the hands reveal about the brain*. Cambridge, MA: MIT Press.
Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In D. Allport (Ed.), *Language perception and production: Relationships between listening, speaking, reading and writing* (pp. 74). London: Academic Press.
Volterra, V., & Erting, C. (1990). *From gesture to language in hearing and deaf children*. Berlin: Springer-Verlag. pp. 302–303.