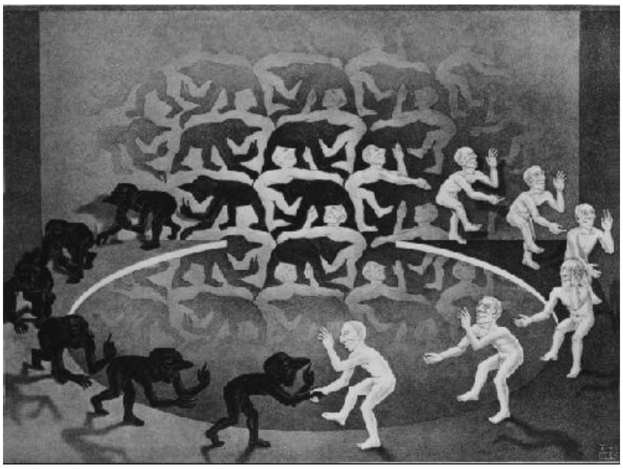
Action representations and the semantics of verbs



Matthijs Westera

Bachelor's Thesis

June 2008

Cognitive Artificial Intelligence
Utrecht University

Supervisor:
Dr. J. Zwarts
Linguistics, Department of Foreign Languages
Faculty of Humanities
Utrecht University

Front page illustration: Encounter, by M.C. Escher, 1944.

I have always thought the actions of men the best interpreters of their thoughts.

- John Locke in *An Essay Concerning Human Understanding*, ch.2 (1689).

Abstract. This thesis investigates how the syntactical behavior as well as the semantics of verbs can be grounded in the cognitive architecture responsible for action representations. Psychological research on action representation in the human brain is discussed. Based on this, a composite conceptual space for action representation is formalized to model as closely as possible human action representation, including both motor patterns and intentions. A modular conceptual space logic is introduced to denote action concepts in a natural way. Using this logic, the manner-result dichotomy in verb meaning, the aspectual structure in event descriptions and the special status of modal verbs are shown to be explicable in terms of the model described. Throughout the thesis, the computability of the model is taken into account for future implementation, e.g. in robotics. Various valuable suggestions for further research are given.

Keywords: conceptual space, action representation, motor pattern, intentionality, verb meaning, aspect, cognitive semantics.

0 Table of contents

1 Introduction[†]

The information density in dynamic scenes is very high. Consider the straightforward scene of a smith hitting a piece of heated metal with a large hammer. When investigating the scene we could focus on the hammer hitting the metal, on the metal changing shape, on the smith's arm swinging, on the smith's will power to continue the work regardless of his blisters, on the smith's intentions to earn money by creating a neat sword for his landlord, on the causal chain initiated by the muscle contraction in his right shoulder... and so on.

Fortunately the human brain has found ways to discriminate between and cope with many of these various aspects of a dynamic scene, and this ability is reflected in our language. We have different types of verbs for describing the manner and the result of an action, compare "hit" with "break", "pour" with "fill", "stab" with "kill". Lexical aspect in turn seems to discern between various basic types of events such as states, activities, achievements and accomplishments, which can in turn be divided into more specific subcategories. On a higher level we have words like "despite", "cause" and "prevent" to describe the causal structure of a scene.

1.1 Objectives

In this thesis I attempt to show how structures in natural language are rooted in our cognitive architecture. The focus is on the representation of physical actions involving a human agent and possibly a non-sentient patient and one or more instruments. For this purpose a framework will be formalized for the representation of such actions that resembles the responsible faculty or faculties in the human brain. In this I pursue the conceptual spaces approach advocated in (Gärdenfors 2000). Insights from neurology and evolutionary biology will be used to formalize a conceptual space for the representation of actions. When a solid framework is in place I will investigate how three linguistic phenomena could stem from this kind of action representation: the manner/result dichotomy in verb meaning, the event types found in the analysis of lexical aspect (e.g. states and activities) and the verb category of modals (e.g. "can", "may", "should", "ought").

[†] I wish to thank my supervisor Joost Zwarts for his valuable suggestions, for his detailed, inspiring comments, and in general for the fruitful conversations on the subject.

My study, Cognitive Artificial Intelligence, which is highly multidisciplinary, aims to combine insights from cognitive science with the field of artificial intelligence. This cross-fertilization is reflected in my thesis: I aim to construct the action representation framework in a way that is fairly directly applicable in robotics. For that reason, considerations from the field of artificial intelligence and robotics will appear here and there throughout this thesis.

1.2 Relevance

Actions are central to human behavior and cognition. Not only are they the basic ingredients of human behavior, they also seem to play an important role in higher cognitive functions and in the classification of objects by means of their affordances. This thesis contributes to the study of human behavior and cognition by summarizing the existing neuropsychological data on action representations in a formal model. Obviously, these data are incomplete. For that reason some plausible assumptions are made based on the evidence available and several hypotheses are raised. This thesis thus contributes to the field also by providing a good guideline for further research. Additionally, several linguistic phenomena, which are rarely related to each other in the literature, are accounted for by looking at action representations. The model in this thesis can be seen as a first step towards a unifying framework in cognitive semantics. The model also makes several empirically testable predictions in this area.

More generally, this thesis provides insight in the nature of concepts, which has been subject of debate for a long time. Frameworks like the classical definitional theory and its successor, prototype theory, have come and gone. The conceptual spaces framework advocated by Gärdenfors (2000) is yet another attempt to capture the true nature of concepts. It seems to be an empirical issue whether or not these conceptual spaces are really expressive enough and are fit to the job of concept modeling. By attempting to formulate a conceptual space for the representation of actions, knowledge is gained on the plausibility of the conceptual spaces framework in general.

This thesis also has some more practical contributions. For artificial agents to be truly intelligent, they have to be able to plan their own actions as well as interpret the actions and intentions of other agents, including humans. In general, for human-machine interaction based on visual information a good model of human actions is required. The framework presented in this thesis will serve as a good baseline model that is in principle readily implementable. Similar but less expressive frameworks have already shown to be applicable for the classification of dynamic scenes, e.g. (Chella, *et al.* 2000), and in imitation learning by robots, e.g. (Ilg, *et al*. 2003).

1.3 Structure

This thesis is structured as follows. Chapter 2 concerns the general nature of action representations. An overview is given of existing action representation frameworks (symbolic, associationist and conceptual approaches). From evolutionary and neurological considerations I will conclude that action representations should contain a description of the motor patterns and of the goal of the action.

In chapter 3 a conceptual space for the representation of actions is formalized. I will first introduce the conceptual spaces approach in general and define a modular conceptual space logic. Then a conceptual space for the representation of motor patterns and another for the representation of goals are formalized. Both will be used for the representation of actions in what I will call *action space*.

In chapter 4 I will argue how three important linguistic phenomena stem from our cognitive architecture, represented by a composite conceptual space. These phenomena are the manner-result distinction in verb semantics, lexical aspect and its event types, and the syntactic category of modals.

In chapter 5 I will give a brief summary of the framework and my findings and draw a number of conclusions. Several interesting suggestions for further research are given.

2 Towards action representations

This chapter discusses the general nature of action representations. In section 2.1 I will provide an overview of existing action representation frameworks (symbolic, associationist and conceptual approaches). After that in section 2.2 I will argue from a neurological and evolutionary perspective that action representations should contain at least a description of the motor patterns involved and of the goal of the action. Section 2.3 provides a short summary.

2.1 Previous approaches to action representation

The first approaches towards action representations were highly *symbolic*, based on various types of logic. Originally, actions were represented as abstract collections of high-level facts about conditions, effects and goals, without caring about the motion patterns that constitute actions in the outside world. From the eighties onwards a different, but also symbolic, approach to action representation was seen. In these models the description of motion rather than intentions, conditions and effects was considered an important part of action representations. These methods usually decompose a dynamic scene in formal descriptions of geometric shapes and motion patterns.

Complementary to the symbolic approaches is *associationism*. In associationist models, concepts (e.g. actions) are represented usually in a network of associations with other concepts. A popular branch of associationism is *connectionism*, which uses artificial neural networks, consisting of a large number of interconnected artificial neurons. In the context of action representation, connectionist networks are usually used to categorize motion patterns, ignoring any higher-order properties of actions.

In between the symbolic and the associationist frameworks are *conceptual* approaches which are concerned with the representations of meanings at an intermediate level. The constructs in these approaches, generally called concepts, are fairly directly linked to perception, although usually some preprocessing on the perceived data is done. The (preprocessed) data is ordered in some sensible way, e.g. in a metric space.

In this section I will give examples of existing approaches in each of the areas described above: symbolic, associationist (connectionist) and conceptual. I will also discuss which aspects of each approach are useful for the objectives of this thesis, and which aspects are not. Note that it has been argued, among others by Gärdenfors (2000), that the three approaches should not be seen as mutually exclusive frameworks. Instead each has its own advantages and disadvantages and in the end hybrid approaches may turn out to be the most effective. Note also that, even though I explicitly discern here between the three kinds of approaches, Polk *et al*. (2002) propose that there is a natural mapping from symbolic, goal-driven cognition onto connectionist models. Similarly, Chella *et al*. (2001) propose a mapping between *conceptual spaces*, the conceptual approach that will be used in this thesis, and *situation calculus*, a symbolic approach. However, the existence of such mappings does not mean the approaches are the same 'in principle'. The different approaches all have their own merits and disadvantages. Also, the two mappings mentioned are only fit to very specific instances of the three approaches and are not easily made more general.

2.1.1 Symbolic approaches

A widely used symbolic approach to high-level action representation is a logic called *Situation Calculus*, first treated extensively by McCarthy and Hayes (1969). Situation Calculus is a first-order logic attempt to represent actions and their effects on the world. The building blocks are actions that can be performed, fluents that describe the state of the world and situations that occur. A domain is described by a set of formulae, among which are those that determine the preconditions for actions. An example if given in (2.1), which says that it is only possible to drop something in a situation s when you are actually carrying it (this and the next example are taken from Wikipedia).

$$Poss(drop(o), s) \leftrightarrow is\_carrying(o, s) \tag{2.1}$$

In addition to these preconditions, the effects of actions need to be specified. I will not go into the problems concerning the formulation of action effects (and non-effects), collectively referred to as the *frame problem*, see (McCarthy and Hayes 1969). One step towards a solution is to enumerate as *effect axioms* all the ways in which a fluent (e.g. the broken-ness of an object) can be changed as the result of an action, e.g. (2.2).

$$Poss(a, s) \rightarrow \begin{bmatrix} broken(o, do(a, s)) \leftrightarrow & \\ & a = drop(o) \wedge fragile(o) \ \vee \\ & broken(o, s) \wedge a \neq repair(o) \end{bmatrix} \tag{2.2}$$

In addition to these high-level reasoning formalizations, several attempts were made at formulating geometrical models of the human body to represent motion as opposed to actions and their effects. A main problem in such approaches is posed by the need to segment a continuous stream of movement into elementary chunks of motion that can be represented. A framework that does this in an adequate way was proposed fairly early by Marr and Vaina

(1982). Their framework is grounded in a cylindrical model of the human body. Body parts are represented by cylinders and an intricate coordinate system allows for the representation of the movement of these cylinders with respect to one another. An important aspect of the model is that it can be looked upon at several levels of granularity. This is displayed in Figure 1, which depicts a walking person at different levels.

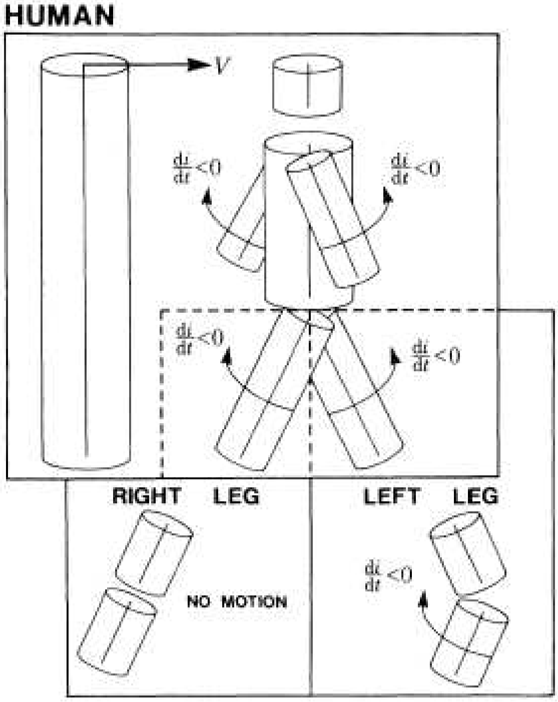


Figure 1. An illustration of the geometrical framework described in (Marr and Vaina 1982). Taken from (Marr and Vaina 1982).

To be able to find the most elementary chunks of motion they exploit the fact that in any motion pattern there is a moment in which limbs are at absolute or relative rest, e.g. for walking there are intuitively two ‘states’, with the legs at the extremities, and two ‘motion segments’ with the legs swinging. They thus introduce the *state-motion-state* decomposition of motion. I will skip the formalities, but it is important to note that at each level of granularity there is a different decomposition. At the most global level in Figure 1 there is in fact no state at all, just motion. Looking at the motion in some more detail, there is per step a clear state-motion-state decomposition. Going to an even more fine-grained level, zooming in on the legs, the motion parts resulting from the first decomposition are themselves decomposed again. Although it is a highly simplified model and no explicit attempt has been made to add biological realism, it is very robust and in fact quite expressive. The state-motion-state decomposition, although it may seem naive, has been adopted in a slightly different fashion by more recent (both associationist and conceptual) approaches, e.g. (Chella, *et al.* 2000), (Giese and Poggio 2002).

Several general objectives apply to symbolic approaches to knowledge representation, as formulated for example in (Gärdenfors 2000). One that is mainly important for the aims of this thesis is the *symbol grounding problem*. Symbols do not have true meaning, in principle. One symbol can be defined in terms of others, but you will always end up in loops or an infinite regression trying to find meaning primitives. Harnad (1990, 2005) discusses this problem and argues for the embodiment of artificial intelligence to ground the meaning of symbols in sensory-motor experience. The problem however is that symbols are not naturally grounded in perception. There are no symbols *out there*. Instead, the world is a continuous, fuzzy process that has no clear boundaries. Especially for the understanding of human cognition a symbolic approach seems inappropriate.

2.1.2 Associationist approaches

An associationist, connectionist approach is described in (Giese and Poggio 2002). Their main aim is to summarize in one artificial neural network most results from (neuro)psychological research on action recognition and representation. The model consists of a form pathway and a motion pathway that are given the same motion scene as input. Each pathway consists of a number of stages, each stage extracting features of higher complexity. For example, the form pathway consists of first some simple basic filters, then cells that respond to specific orientations, then neurons selective for entire body poses and finally neurons that represent to actual motion patterns. The processed results from the two pathways are merged in the end. Interestingly, the model makes use of encoded ‘snapshots’ from the movement that are strikingly similar to states in the state-motion-state decomposition of Marr and Vaina (1982):

> “The neurons at the next level of the form pathway are functionally equivalent to the “view-tuned neurons” that have been found in area IT of macaques. [...] After learning, such neurons encode “snapshots” from image sequences showing complex body movements.” (Giese and Poggio 2002)

A problem with connectionist (and most associationist) models of cognition in general is that they tend to give rise to *Bonini’s paradox*, which has been articulated in a modern sense as follows:

> "As a model of a complex system becomes more complete, it becomes less understandable. Alternatively, as a model grows more realistic, it also becomes just as difficult to understand as the real-world processes it represents." (Dutton and Starbuck 1971).

This problem does not have much of a role in computer science, where biological realism is not strived for and neural networks are generally employed simply as very efficient pattern matchers. But also for modeling human cognition Bonini's paradox is no reason to immediately discard artificial neural networks as a candidate. The virtue of artificial neural networks for cognitive modeling, as opposed to their biological counterparts, is that once the network has been trained the researcher can freely wreak havoc on it and slowly tear it down into to pieces, keeping track of the functionality loss. Also, it is very easy in an artificial neural network to 'tap a wire' (e.g. track single-neuron behavior), which is done sometimes, but with much less precision, during open brain surgery. So the easier pruning and wire tapping may push Bonini's paradox to the background. Medler and Dawson (1998) explain that appropriate statistical analyses can and do shed light on the algorithms encoded in neural networks, even to the extent that they have greater explanatory powers than classical (e.g. symbolic) models.

The need to analyze a trained network afterwards may be a burden. There is however also a practical objection to connectionism. A neural network that is only moderate in size and expressivity already needs a very large amount of data to be able to successfully learn in the first place. Such data are, in the domain of actions, currently unavailable. An associationist approach, although very successful in various areas of computer science, does not suit the objectives and methodology of this thesis very well. Fortunately there is one more alternative.

2.1.3 Conceptual approaches

One of the advocates of a conceptual approach is Peter Gärdenfors. He proposes a general theory of concepts based on geometrical structures in so-called *conceptual spaces* that are defined by quality dimensions (Gärdenfors 2000). The ins and outs of the conceptual spaces approach will be discussed later in chapter 3. Nevertheless I will discuss several conceptual spaces approaches already, skipping the details.

In (Gärdenfors 2007) a first hint is given towards a conceptual space for action representation. Gärdenfors argues that actions should be represented as spatio-temporal patterns of forces – a key notion that will appear later in this chapter. He takes on an 'embodied' perspective, arguing that the most important forces are the ones that act on the agent itself. Furthermore, he suggests that functional properties (i.e. affordances) are convex regions in action space. This would explain the role of affordances in the visual recognition of objects, especially instruments, as discussed by Helbig and his colleagues (2002). Even though Gärdenfors' account of action space remains rather abstract – how should those forces be represented? – he gives a great overview of the evidence in favor of an approach along these lines.

Geuder and Weisgerber (2002) take on a less abstract approach and explore possible ways to adapt the conceptual spaces to suit verb meanings. They introduce a conceptual space to account for the semantics of “verbs of vertical movement”. Through various examples they illustrate that two main features seem to be the direction and the manner of the movement. The resulting two-dimensional space is given in Figure 2, with regions corresponding to the semantics of several German verbs.

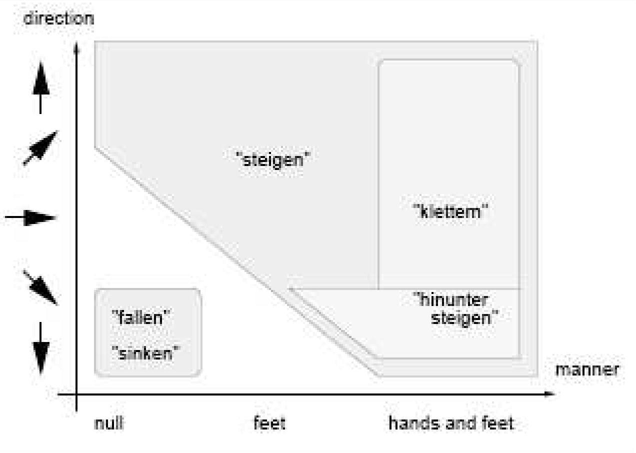


Figure 2. A conceptual space for verbs of vertical movement, with regions corresponding to the meanings of various German verbs. Taken from (Geuder and Weisgerber 2002).

Geuder and Weisgerber note that this simple conceptual space cannot account for the difference between ‘fall’ and ‘sink’ (compare “Das U-Boot sinkt tiefer” with the ungrammatical ”*Das U-Boot fällt tiefer”), so they propose the addition of two features, the medium in which the movement occurs and the attachment to e.g. an object, wall or floor, that are in fact not independent dimensions, but different values on the same dimension. They further note that several verbs are *polysemous*, with a meaning fully dependent on these two attributes, and therefore suggest two distinct conceptual spaces to capture this.

Geuder and Weisgerber’s approach is certainly insightful but it is very specific and seems to lack the expressivity that I need for the aims of this thesis. In fact they extrapolate the specificity of their framework to the general case and conclude the following:

> “Employing conceptual spaces in order to give a lexical analysis [...] actually has to be viewed as nothing more than an attempt to construct a particular conceptual space that optimally serves a particular purpose. What the “lexical representation” did was to try to establish a particular grouping of items (by way of a particular selection of feature dimensions) so as to maximize

> comparability – at the expense of broadness of coverage. But maybe this is a deeper point about lexical representations, which is not to blame on the geometric model: that conceptual features do not exist in isolation but can only be established in the course of comparison operations." (Geuder and Weisgerber 2002)

Geuder and Weisgerber seem to ignore the fact that these features are not transcendental Ideas to which humans obey, but are a natural result of the human perceptual apparatus. What they should have done, to avoid the specificity of their conceptual space, is formulate a more general action space and see how such features as direction, manner and medium flow naturally from it. (Indeed, it will become apparent that the features direction and manner are accounted for very directly in the framework that is formalized in chapter 3, and that also a feature like medium can be retrieved from it).

A more expressive approach is presented by Chella, Frixione and Gaglio (2000, 2001). They propose a framework that bridges the gap between low-level motion recognition systems and high-level, symbolic representations of processes and actions. A conceptual space is employed to bridge this gap between on one side what they call the sensory data and subconceptual area and on the other side the linguistic area. In the subconceptual area, a geometric description of the dynamic scene is generated. In the conceptual area this description is represented in chunks of information that correspond to 'simple motion', similar in meaning to 'motion segment' in the approach of (Marr and Vaina 1982) described above. The simple motion of an object is represented as a vector in a metric (conceptual) space, which they call a *knoxel*. In Figure 3 this conceptual space is illustrated below the dashed line, with groups of knoxels corresponding to a type of motion (e.g. upper arm stretching).

So, below the dashed line in the figure is the conceptual area, where regions in the conceptual space correspond to certain motion patterns. Above the dashed line is the linguistic area, where the actual expressions dwell and are combined into expressions with a more composite meaning such as "arm approach" (which is a combination of upper and lower arm stretching) and ultimately expressions describing more complex actions such as "seize". The name "linguistic area" suggests a large correspondence of the constructs in this area and linguistic expressions. However, it is just a level of higher-order conceptual structures, some of which indeed may accidentally correspond to verbs like "seize" and "grasp". Most of the constructs in the linguistic area do not have a direct linguistic counterpart.

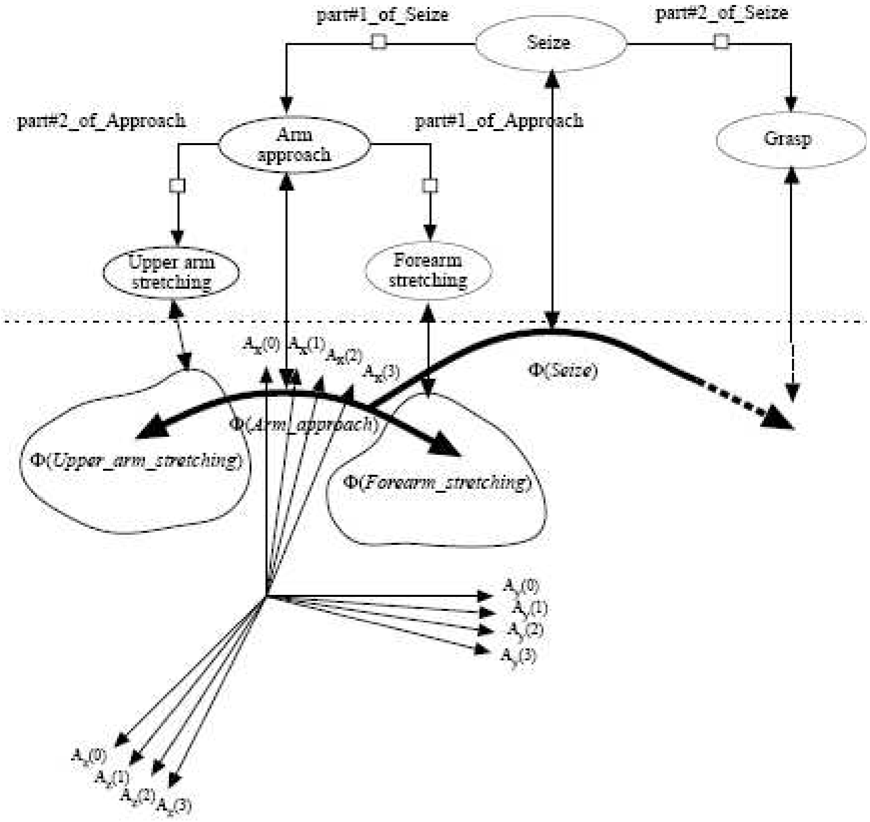


Figure 3. A pictorial representation of the meaning of "seize", taken from (Chella, *et al.* 2000).

There are several points that make this approach very appealing. Firstly, it is grounded in the conceptual spaces approach, which appears to be a very elegant and promising approach to representing human conceptualization. Secondly, the approach is one of the very few conceptual approaches to action representation that is actually formalized and ready to be implemented. Thirdly, a link from the conceptual area to symbolism is established, which is related to the aims of this thesis because natural language is essentially a symbolic system. Fourthly, the framework includes an attention mechanism that adds computational efficiency, but also makes sense from a cognitive perspective. The attention mechanism utilizes linguistic expressions, e.g. "grasp", to focus on a particular area of the conceptual space, looking for shapes and motion that matches the expectation. I will not present such an attention mechanism in this thesis, however for applications in artificial intelligence, adding one to the framework formulated in chapter 3 will be very useful.

However, there are also aspects that make this approach less attractive for the aims of this thesis, which require cognitive realism, but also for applications in artificial intelligence, mainly due to the lack of expressivity. For instance, the conceptualization relies on motion whereas there are many actions that do not involve motion, such as holding something or pushing against a wall. The human capacity to recognize causality in dynamic scenes is grounded in the our perception of forces (Wolff 2007, Barbey and Wolff 2007), but the notion of force appears nowhere in the framework. Furthermore, it should be noted that the constructs in the linguistic area can only represent motion, while there is more to the semantics of verbs than just motion. An important aspect of actions seems to be the goal or intention of the agent. How else can we discern between the unfortunate bungler who throws a rock and causes a person to die and a willful murderer who does the same thing to kill someone? However, no such thing as intention or goal is included in their approach. Both the importance of forces and of goals will be argued for below.

In summary, a conceptual approach like the Gärdenforsian conceptual spaces framework suits the objectives of this thesis the most. The ideas of Gärdenfors on action space combined with the attempt by Chella and his colleagues provide a great starting point.

2.2 Action representation ingredients

There are neurons in the brain that display activity both when performing an action and when *observing* someone else performing the same action (Kilner, *et al*. 2004). So, not only do we recognize actions, we also experience observed actions in the same way as we experience our own. Interestingly, the same mechanism has been indicated in the monkey brain (mirror neurons) by, amongst others, (Rizzolatti, *et al.* (1996), Kohler, *et al*. 2002). Independently, Feldman and Narayanan (2004) conclude that the relevant areas in the brain do not only serve the control of action, but also serve the construction of *representations* of actions, including the objects acted on and the locations towards which is acted. So, it seems that a good way to learn about action representation is to study both the recognition and the generation of actions.

Two things seem to contribute to the recognition of an action. On one side, we recognize the motor patterns. Note that motor patterns need not cause actual motion, since, as mentioned, there are certain static actions that result from complex motor patterns. On the other side, we almost instantly start interpreting and recognize the intentions of the agent performing an action. This uniquely human capacity is often looked upon as a key step in the evolution towards real symbolic language and thought, and has been discussed in depth amongst others by Tomasello and his colleagues (for example 1997, 2005). So both motor patterns and the agent's goal contribute to the recognition of an action.

The same division is present in the generation of actions. In (MacKay 1987) a hierarchical model of motor control is presented. At the top level, the *conceptual level*, the goal of the action is generated. At the lowest level, the *motor implementation level*, the actual motor plan is generated and finally impulses are sent to the muscles. An intermediate level, the *response system level*, selects the appropriate response to achieve the goal. This hierarchy reflects how actions may be learnt or improved. When learning an entirely new action, learning occurs mostly on the conceptual level. When performing an action in a different manner than usual (e.g. left-handed instead of right-handed), learning occurs mainly on the motor implementation level.

In both the recognition and generation of actions, there seem to be two main ingredients: a motor description and a goal description. Accordingly, this section is twofold. The first part argues, following (Gärdenfors 2007), for the importance of forces and discusses the role of effector systems in the representation of the motor patterns involved in actions. The second part argues for the importance of goal-directedness in action representations and elaborates on goal representation in humans. Hints towards the formalization in chapter 3 will be given throughout this section.

2.2.1 Motor patterns

Motion can be represented in several ways. It can be represented as a series of strokes for certain limbs with a certain velocity. It can be represented as a series of snapshots, showing several stances of the body and leaving the actual motion to be inferred. However, both of these approaches lack expressivity and biological plausibility. In this section I will argue for a third approach, which is based on forces, following (Gärdenfors 2007). This choice makes sense intuitively: no velocity-based representation could appropriately represent actions that involve no motion.

A more convincing argument for the modeling of human action representation with force patterns can be derived from Wolff's work on causation (2007), which investigates the building blocks of human causal judgments. Wolff discerns between *dependency models*, in which causal relationships are represented as (possibly probabilistic) contingencies between causes and effects, and *physicalist models*, which tend to model causation in terms of physics (e.g. momentum). Wolff argues that dependency models do not model human cognition appropriately, and instead he argues for a physicalist approach, in particular one based on force dynamics, his *dynamics model*. Wolff argues for the patient tendency P, the affector force A and other forces O as the basic ingredients of a dynamic scene, as depicted in Figure 4. These forces, together with the resultant vector R and the endstate vector E, form the building blocks of causal judgments.

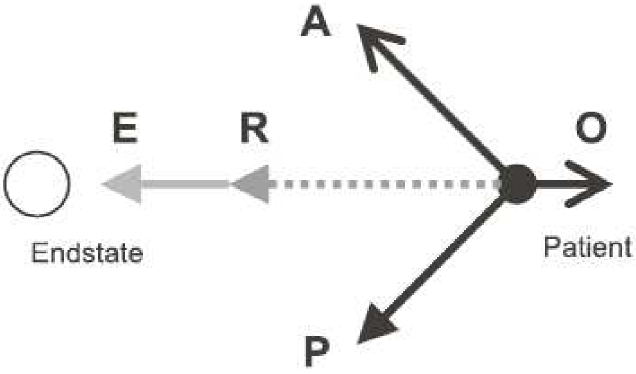


Figure 4. The basic ingredients of the dynamics model. Taken from (Wolff 2007).

Wolff predicted that our use of the words “cause”, “enable”, “prevent” and “despite” is based on three Boolean features: the collinearity of patient tendency P and endstate vector E, the collinearity of affector force A and patient tendency P, and the collinearity of resultant R and endstate E (e.g. “cause” is predicted for the configuration $\langle no, no, yes \rangle$). To test these predictions, Wolff generated several scenes in a physics simulator, varying the magnitudes and directions of the vectors. Participants were then asked to describe each scene with any one of the words “cause”, “enable”, “prevent” and “despite”. The dynamics model, which is based on forces, gave very accurate predictions on human causal judgments. In addition, as is shown in (Barbey and Wolff 2007), this force dynamical approach can account for causal reasoning, e.g. inferring “vegetation prevents landslides” from “vegetation prevents erosion” and “erosion causes landslides”.

Gärdenfors emphasizes that the forces represented by the brain are not the Newtonian entities, but psychological constructs that could be called “powers” to avoid the physical connotation. Though this is certainly true, in this thesis it is more convenient to ignore it and model Newtonian forces nevertheless. The main reason is that little is known on the precise nature of folk-physics, whereas Newton has provides us with a beautiful formalization that can be modeled easily. This can be done whilst preserving cognitive realism, because the psychological constructs of folk-physics can be mapped onto the Newtonian entities. Regardless of whether this mapping is a true isomorphism (probably not), it must be reliable enough to allow us to move around and to manipulate objects and will be largely similarity-preserving. This means that there is no significant loss of information – as far as moving around and manipulating objects are concerned – when going from the Newtonian framework to folk-physics, and vice versa. Besides this, Gärdenfors hints at the inclusion of psychosocial forces, but these are outside the scope of this thesis.

The resultant force can be decomposed into its various components, among which are the force applied by the agent, gravity and friction, and a choice has to be made concerning which components to include in the motor representation. At the motor implementation level, a representation of the force applied by the agent's muscles is required, because otherwise no appropriate signal can be sent to the muscles. Intuitively, the other forces as well as the resultant force have more to do with the goal of an action than with the motor representation. To lift an object, an upward resultant force has to be exerted on the object. This goal is formulated somewhere at a higher level, and the weight of the object is subtracted from this intended resultant force to obtain the right signals to be sent to the muscles. Little is known on the subject, so I will just follow this line of thought and only include the agent forces in the representation. I will address the representation of the goal further below.

So how are forces extracted from perceived data? Gärdenfors points at research done by Runesson and Frykholm (1981) on motion recognition from joint positions alone, using patch-lights attached to a human dressed in black. Subjects were shown to be able to extract subtle details of the action, such as the presence of unnatural weights. I follow (Gärdenfors 2007) by including the same quote he does:

> "The fact is that we can see the weight of an object handled by a person. The fundamental reason we are able to do so is exactly the same as for seeing the size and shape of the person's nose or the color of his shirt in normal illumination, namely that information about all these properties is available in the optic array." (Runesson 1994, pp. 386-387)

I agree with the implications of this statement, namely that force extraction from perceived scenes happens automatically and instantly, but the statement itself is arguably too strong. Extracting forces from perceived data is a complex process. According to equation (2.3) what is required for the extraction of (Newtonian) forces is the first order derivative of speed, i.e. the acceleration, and the mass of the object.

$$\vec{F}_{res} = m \cdot \frac{d\vec{v}}{dt} \tag{2.3}$$

Estimating the mass m of an object seems to require some basic knowledge of the world, such as rules of thumb that infer the weight from the perceived size and possibly some knowledge of the materials involved. This kind of indirect perception is of course prone to illusions. Someone can pretend that something is really heavy and then lift it up easily to give an illusion of super strength. So, obtaining the resultant force $\overrightarrow{F}_{res}$ from a perceived scene is already non-trivial. But it gets worse.

Even though the resultant force determines whether an action is successful, as I discussed above it is not the force that should be in the motor representation. It has to be decomposed somehow into its component forces. Unfortunately there is no way to unambiguously decompose a resultant force into its components. So, while 'seeing' the forces seems to happen automatically, it is certainly not as trivial as seeing colors or shapes, as suggested by Runesson. In fact, in (Povinelli 2000) it is shown that non-human primates lack such a capacity entirely: they often fail to understand the hidden forces behind effects. Fortunately, the main aim of this thesis is not endangered by the difficulty of force extraction. It poses a problem only for the side quest of this thesis, which concerns the applicability of the action representation in robotics, and therefore I will suggest the following approach to tackle this problem. What seems to be the case is that, to achieve the decomposition of a resultant force, we make use of our knowledge of natural *tendencies* of objects. Trees usually stand upright, and any different orientation may lead us to believe that there are some strange forces at work (e.g. the wind or a dangerous monster). Analogously we know from our own experience what stances of the human body are the most efficient, so if a human body appears to be in a stance less efficient or comfortable, then the presence of an unnatural weight or other kind of force is inferred. For that reason I think an appropriate set of heuristics related to this kind of tendencies and world knowledge can achieve a plausible force decomposition. Alternatively, an artificial neural network may very well be trained to do the job, judging from its success in other visual classification tasks. I will not go any deeper into this.

As mentioned in the previous section, state-motion-state decomposition as advocated by Marr and Vaina (1982) is based on the occurrence of states of relative or absolute rest in movement. This allows perceived motion to be segmented with relative ease, into what are called motion segments in (Marr and Vaina 1982) or simple motions in (Chella, *et al.* 2000). But where do these states of relative or absolute rest occur in a force dynamical representation? States are moments with zero velocity, which are often points in time at which the direction of the velocity suddenly reverses. These changes require a very high deceleration/acceleration. For example, in a walking motion, the state of rest between two steps co-occurs with very high forces (because the legs decelerate/accelerate very fast).

So a consequence of the force dynamical approach is that the state-motion-state decomposition as proposed in (Marr and Vaina 1982) and recycled in (Chella, *et al.* 2000) is no longer applicable. This, however, is not a problem for the aims of this thesis. Sternad and Schaal (1999) show that the apparent segmentation of trajectories (velocity-wise) does not imply segmented motor control, so the classical state-motion-state decomposition is not necessarily present in human action representations. Nevertheless it will be convenient to somehow cut the continuous stream of motion into representable blocks, and fortunately an alternative to the state-motion-state decomposition is available. *Motor segments* could be defined as intervals of motor control in which the direction of the force does not change significantly, i.e. intervals during which the neuronal signals sent to the muscles do not change. Such force segments are separated by sudden shifts in force, e.g. when the smith's hammer hits the iron and when the legs of a walking person switch (halfway) from acceleration to deceleration and vice versa. This will become more apparent in chapter 3.

Interactions with the outside world always happen via an *effector system*, e.g. the human hand together with all the joints controlling its motion (wrist, elbow and shoulder). Effector systems thus play an important role in actions. There is a peculiarity with these effector systems. When learning to perform the same action with a different effector system, e.g. left foot instead of right hand, an entirely new motor plan need not be created. Instead, the existing motor plans are easily adjusted to different effector systems. This process is called *transfer*. Figure 5 shows the results of performing the same action (writing) with five different effector systems.

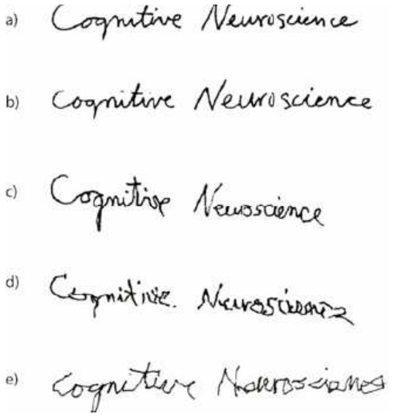

a) Cognitive Neuroscience
b) Cognitive Neuroscience
c) Cognitive Neuroscience
d) Cognitive Neuroscience
e) Cognitive Neuroscience

Figure 5. A demonstration of effector-independent motor representation. These five productions of "Cognitive Neuroscience" were written by moving a pen with a) the right hand, b) the right wrist, c) the left hand, d) the mouth and e) the right foot. From (Gazzaniga, Ivry and Mangun 2002).

Disregarding the differences in smoothness that are due to a lack of practice with some effector systems, the apparent similarities in the results show that motor representations are somehow independent of the effector system.

This effector independency seems to reflect our intuitions. When someone is holding a pen between his toes and clumsily attempts to write his name on a sheet of paper, this is intuitively still called "writing". Holding a club between your teeth and swinging it can still be called "hitting" without much confusion. Ideally, an action representation framework should be expressive enough to account for such effector independency. On the other hand, some actions are impossible to achieve through any effector system but the regular one, due to human anatomy. But this kind of overexpressivity is no problem for an action representation framework – some action-effector combinations just never occur. I will come back to this in chapter 3.

2.2.2 Goals

Goals are an important part of actions. To clarify what is meant by goals *in this thesis*, consider the following tentative question. Do humans, or living creatures in general, *ever* perform an action *without* a goal? Surely some actions are done automatically, you might say. But even these subconscious movements are directed towards a goal somehow, whether or not this goal is itself subconscious. A subconscious scratching of the nose is aimed at taking away the itching and it may or may not succeed at that. Just as it makes no sense to deny subconscious actions a motor representation – there is motor control, after all – it is unreasonable to deny such actions a goal representation. Then how about wandering around aimlessly? The lack of a high-level, cognitively established goal does not entail the complete lack of lower-level motoric goals. In such cases, motion itself can be considered the goal of the motion, just as the goal of a crawling worm is to crawl. An argument for the existence and significance of such low-level motion-goals is that an action can only succeed or fail when the action was aimed at something. Surely wandering around aimlessly can fail, for example when the person is interrupted and asked to do the dishes or, if the wanderer is really persistent and keeps on wandering, at least when the wanderer twists an ankle. I cannot think of any action that cannot succeed or fail, nor of actions that could not be slightly more or slightly less successful (as a last resort, the agent could always die in order to fail, unless if the action was to commit suicide). In the sense of goal *employed here*, every action has a goal. It requires no conscious intentionality or high-level aims.

This short philosophical enquiry may or may not be convincing, at least it assisted in clarifying the meaning of ‘goal’ in this thesis. What matters is that the action representation framework I propose in this thesis will allow goals of the type discussed here. If someone decides to use my framework but wishes to differentiate further between what can and what cannot be called ‘real goals’, or ‘conscious goals’ if you will, extra constraints could be imposed.

The emphasis on the inclusion of low-level motoric goals does not do justice to the amazing human capacities for higher-level goal recognition, goal formation and planning, that together with our language and symbolic reasoning capacities are an important cause of the cognitive canyon between *homo sapiens* and their nearest primate relatives, see for example the research done by Tomasello and Call (1997) or Suddendorf and Corballis (1997). Whereas some apes can solve problems quite successfully with the objects available, such as the stacking of crates or the use of a stick to reach a banana in a high place, they are helpless when the instruments are not in the immediate surroundings of the problem scene, even when these apes have the knowledge that the perfect instruments are just around the corner (one of the first to notice this was Köhler in 1925). Humans, on the other hand, actively search for the instruments they need, also when they do not know they are just around the corner, contrary to the ape scenario. Most apes can only plan with the affordances of the objects present, whereas humans actively look for the means to reach the ends. This is the difference between *reactive planning* on one side and *means-ends planning* on the other. There is no agreement on whether means-ends planning is simply an enhanced version of reactive planning or whether it contains a truly new mechanism. In any case it is obvious that this capacity is paired with a much more rich and powerful goal representation than the goals found in non-human animals.

In the section above on motor representation I briefly looked at the mechanism responsible for the perception of forces. In the same way I will here devote a paragraph to the recognition of goals. Just like we appear to see the forces that act immediately, we also immediately start mind-reading, interpreting, trying to understand the goal of the person performing an action. The process of goal recognition is so innate to us and so intertwined to other, broader mental capacities, that little is known about the underlying cognitive mechanisms and opinions diverge about its possible implementation in artificial intelligence. In (Hernandez Cruz 1998) two approaches are discussed. On one hand, the *theory-theory* claims that human adults have a theory of mind (axioms and inference style) that enables them to infer other people's intentions. On the other hand, the *simulation-theory* claims that humans can infer other people's intentions by mentally placing themselves in the same situation and evaluating their own mental states. Hernandez Cruz claims that, in viewing the human brain as a connectionist engine, theory-theories can be implemented only effectively and plausibly in a way that shows in fact many properties of simulation, and thus that the simulation approach should be favored. This would be consistent with the findings of Kilner, *et al*. (2004) discussed earlier, on experiencing the actions of other agents. I will not go any deeper into this issue here, my final remark being that the current evidence seems to hint at a simulation-based approach and that it could be fertile to look at connectionist models for goal recognition in artificial intelligence.

From early research on spatial planning it was concluded that goals are generally represented as a location (i.e. as an endstate). This makes sense computationally for simple actions in which the trajectory is unimportant. However, for most kinds of motion the trajectory can be planned consciously as well. We can change trajectories at will, reproduce movement across a specific distance, and change the speed at which we execute motion. To account for both kind of goal representations a so-called *independent control model* has been suggested (see Gazzaniga, Ivry and Mangun 2002). This model allows an independent location-based and trajectory-based representation of a goal, which may work concurrently, as depicted in Figure 6.

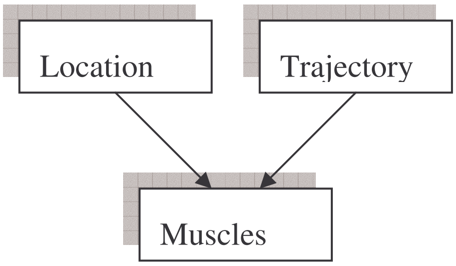


Figure 6. The independent control model for (spatial) motor planning.

The action of pointing is composed of two phases that seem to reflect this model: a rapid movement in the approximate direction (trajectory based) followed by fine-tuning (location based). More suggestive evidence comes from an experiment in (Abrams, *et al.* 1994), where optical illusions are used to affect only one of the two representations.

The independent control model is designed for spatial planning, in which the goals of actions are themselves locations or trajectories. Little research has been done on how this generalizes to more complex actions that are aimed at changing the state in a non-spatial way (e.g. heating, turning, breaking). It seems plausible that a similar distinction can be found there. The goal of putting iron into a fire might be to increase the temperature by a thousand degrees (‘trajectory’ or, more generically, state change) or alternatively to make the temperature equal to a thousand degrees (‘location’ or endstate). The goal of rotating a hammer might be to turn it half a circle (state change) or to turn it into an upside down position (endstate). There even seems to be a difference between hitting to scatter the glass (state change) and hitting to obtain scattered glass (endstate). Even though this generalization of the independent control model to non-spatial tasks has not yet been experimentally confirmed, the twofold representation of spatial goals is well established and the generalization seems plausible. Therefore I will incorporate the independent control model in the formalization in chapter 3. In chapter 4 I will show that this aspect of the model is also linguistically relevant.

Of course, the goal of an action is not restricted to the endstate or state change of a single object. Many actions involve chains of causal interactions. For example, the goal of a smith might be to flatten the iron with a hammer. In such cases, the goal is best represented as a causal chain, as depicted in Figure 7.

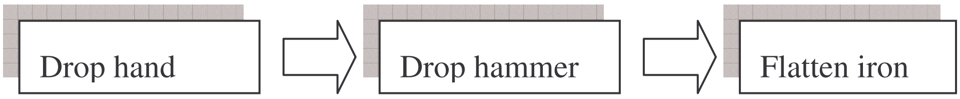


Figure 7. The causal chain of a smith flattening iron.

Such causal chains may be of arbitrary length. The initiator of a causal chain is of course an effector, the human interface with the outside world. A peculiarity of goal representations is shown in (Shen and Alexander 1997b) and concerns the representation of (spatial) goals of actions involving patients, i.e. actions with causal chains as goals. In their experiment an on-screen cursor (the patient) was controlled by the hand movements of the agent. When the direction of the motion of the cursor did not correspond to the hand movements used to steer it, the neural representation of the goal (e.g. move the cursor to the top left corner) correlated with the direction of the cursor motion as opposed to the direction of the hand motion. When the subject got more used to the experiment conditions, the neural activity gradually shifted to a more hand-related pattern.

In more general terms, this would entail that in actions involving a patient, initially the endstate or state change of the patient is mainly represented as a goal. As the agent gets used to the correlation between the agent's own effector system and the patient, the patient disappears from the representation. This *patient-independency* seems a plausible generalization. It would account for the feeling of getting accustomed to an instrument (a hammer or screwdriver, a computer mouse, but also a bike). Whereas a child has to find out how a bike responds to steering and pedaling, professional cyclists are barely aware of the bike. For a cyclist, the goal of steering to the right is represented as leaning to the right, the goal of forward motion is represented as a rhythmical displacement of the feet. Even though this generalization rests in part on speculation, I encourage the reader, in order to be convinced, to introspect the experience of learning to play an instrument, learning to use a new remote control or learning to walk with Nordic walking sticks. This mechanism could be accounted for by assigning different weights to the different parts of a causal chain, which change when learning. This will come back in chapter 3.

Another peculiarity is that goal representations may, at a higher level, be *effector-independent* just like the motor representations discussed above. This is suggested by multiple studies including (Cisek, *et al.* 2003) and (Hoshi and Tanji 2000), that have shown that patterns of activity in premotor regions closely associated with goal representation are largely independent of the choice of the effector system, the left or right arm in the experiments, used to perform the movement. I will return to both peculiarities (patient-independency and effector-independency) in chapter 3.

2.3 Summary

I have given an overview of previous approaches towards action representation and discerned between symbolic, associationist and conceptual approaches. Objections to symbolic and associationist approaches have been risen. Conceptual approaches were shown to be promising and fit for the objectives of this thesis, especially the ones based on the conceptual spaces framework as advocated by Gärdenfors.

I discussed what I consider essential action representation ingredients, from a psychological point of view. Action representations consist of a motor representation and a goal representation. The motor representation is grounded in force interactions, in a way that is at a higher level effector-independent. I have argued for an approach based on Newtonian forces. I have furthermore shown the implications of a force dynamic approach for motion segmentation and argued that the approach raises no unavoidable problems.

Goals may be represented as a state change and as an endstate, according to the independent control model. Goal representation is, like motor representation, at a higher level effector-independent. I have shown that goals can be represented as causal chains involving several instruments or patients and their causal interactions. Furthermore, as a task becomes more accustomed, the representations of the patient and possibly of the instrument disappear from the action representation.

3 A conceptual space for the representation of actions

In chapter 2 I have discussed various aspects of action representations. In this chapter I will formalize an action representation framework based on those considerations. I will take on the conceptual spaces approach advocated by Gärdenfors (2000). Section 3.1 gives an overview of this approach, including the modularity proposed by Geuder and Weisgerber (2005). Also, the syntax and semantics of a conceptual space logic are presented. In section 3.2 a formal account of the more physical part of action representations, motor space, will be presented. Goal space, a conceptual space for the representation of goals, is formalized in section 3.3. In section 3.4, motor space and goal space are combined to form action space and a similarity measure is discussed.

3.1 Conceptual spaces

Conceptual spaces, as advocated in (Gärdenfors 2000), are a framework for representing concepts in a geometrical way. The key merit of the conceptual spaces approach is that it is based on the similarity between concepts, which is, according to Gärdenfors, one of the basics of our cognitive abilities. This section will first introduce the necessary notions for dealing with (modular) conceptual spaces. After that, a logic is defined to deal with concepts.

3.1.1 Quality dimensions, metrics and modularity

Conceptual spaces are constituted by a number of *quality dimensions*. Quality dimensions are the different aspects with respect to which objects may be judged to be similar or dissimilar. Several quality dimensions together constitute a geometric space in which the notion of similarity is represented by a *distance function*, e.g. Euclidean distance. The more similar two points in the space are, the closer they are together. Quality dimensions may be directly related to perception, such as temperature, brightness and the spatial dimensions length, width and depth. However, they may also be of a more abstract character, such as the scale of angriness or happiness, that can only be perceived indirectly. In (Hård and Sivik 1981) a beautiful example of a conceptual space is introduced, though the term “conceptual space” did not exist in the sense it does now: the *Natural Color System*, also known as the color spindle. It is a conceptual space defined by three dimensions: a circular dimension *hue*, a radial dimension *saturation* and a perpendicular dimension *brightness* (or *luminance*). Some slices of the color spindle, each at a different brightness, are shown in Figure 8.

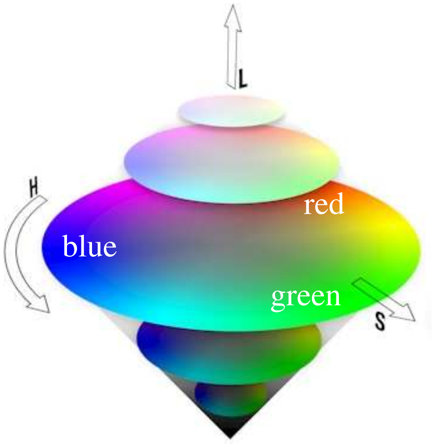


Figure 8. Some slices of the color spindle. Adapted from Wikipedia, author: A. Van de Sande (2005).

Importantly, this color spindle does not model color in the physical sense (which would just be a wavelength/intensity graph), but color as it is perceived by humans. Hård and Sivik's approach is based on the research done by Ewald Hering a century earlier on color opponency in human vision. Color opponency becomes apparent when you investigate the *afterimage* of certain colors. After fixating on a red square for a while, look at a white sheet and you will see a green square, and vice versa. Alternatively, fixate on a blue square, and when you turn away you will see a yellow square. This color-opponency is apparent in the radial *hue* dimension of the color spindle in the figure.

Central to the conceptual spaces framework is the notion of a conceptual *domain*. A domain is a set of related quality dimensions, such as the dimensions that together form color space. Other examples of domains could be shape, position and taste. Regions in these domain spaces correspond to *properties*, like "red", "green", "round", "sweet" and "sour". Categories of objects, such as "apple", can be described very naturally as a collection of properties, e.g. it is red/green, round and sweet/sour. Gärdenfors points in the direction of representing each concept in a high-dimensional conceptual space which combines all the domains that apply to the concept, in a purely geometrical way, e.g. as the Cartesian product. However, Geuder and Weisgerber (2005) notice several disadvantages of this approach. What if we do not know all the properties of an object? Allowing partial vectors, i.e. vectors with some values missing, only partially resolves this, as it is possible that we do not even know which kind of properties apply to an object. Should the conceptual space for the concept "apple" include the property of being organically produced? Also, concepts that reside in different, incompatible conceptual spaces cannot be compared. Unfortunately it is impossible to design one ultra-high-dimensional space in which all concepts reside as partial vectors. Instead, Geuder and

Weisgerber argue for a *modular* approach. Domains are treated as modules that each have a determined inner structure and function as independent, encapsulated data structures. Importantly, domains may be combined *recursively*, allowing composite domains to be the components of even higher-level domains.

In a non-compound conceptual space, suitable distance measures are the Euclidian distance (for entangled quality dimensions) or the Manhattan metric (for independent dimensions). In a complex modular structure of conceptual spaces, an appropriate distance metric is harder to find. Geuder and Weisgerber propose the following similarity measure skeleton, which is some function of the weighted similarities for each domain:

$$\sigma^*(\varphi,\psi) = w_1 \cdot \sigma_{D_1}(\varphi,\psi) \otimes \ldots \otimes w_n \cdot \sigma_{D_n}(\varphi,\psi) \tag{3.1}$$

Here w_i are weights that reflect the fact that domains may be of different importance in a comparison. Weights can be zero to exclude certain domains from the comparison. Obviously exclusion of domains in comparison, especially if they are important ones (e.g. color in the fruit domain), should lead to a lower similarity. (Geuder and Weisgerber 2005) introduce the notion of *comparability* to account for this. It is left open what kind of conjunctive operation $\otimes$ is. I will come back to this at the end of this chapter, in section 3.4.

An important notion in the context of metric spaces is the *convexity* of regions in it. A region in a metric space is called convex if, when two instances X and Y belong to a concept, then all the instances in between X and Y also belong to it. The notion of convexity is illustrated in Figure 9.

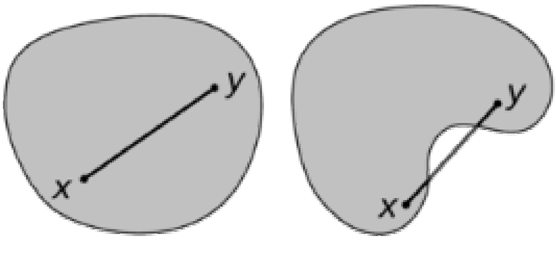


Figure 9. A convex (left) and a non-convex region. From Wikipedia, author: O. Alexandrov (2007).

Gärdenfors (2000) claims that *natural concepts* (informally: concepts that are hard to define in terms of other concepts), are convex regions in a conceptual space. Warglien and Gärdenfors (2007) have argued how the convexity of such concepts makes it possible for agents to agree on a joint meaning even if they start out from different representational meaning spaces. Jäger and Van Rooij (2007) successfully modeled this process in a communication game regarding colors. A division of color space into convex regions was shown to emerge as an evolutionary stable simulation of the game. In this thesis I will only hint at the possible convexity of action concepts.

3.1.2 A conceptual space logic

To be able to speak of quality dimensions, and of points and regions in a conceptual space, in a more formal manner, I will partly adopt the conceptual space logic presented by Fischer Nilsson (1999). However, as I wish to take on the modular approach introduced in (Geuder and Weisgerber 2005), some modifications have to be made. I will first present the logic of Fischer Nilsson, to establish a base, and then present the modifications. In case the reader wishes to skip the formalities and jump to section 3.2, the whole (modified) logic is summarized in the appendix. For the less technical reader mainly the notation conventions in the appendix are worth some attention.

The logic for conceptual spaces is based on concept lattices. The set of possible concepts has a partial ordering $\leq$, which stems from the existence of two operations, lattice join and meet, represented by the operators *concept sum* $(+)$ and *concept crux* $(\times)$. Both operators are axiomatized by the axioms of idempotency, commutativity, associativity and absorption. $[\![\varphi]\!]$ denotes the semantic interpretation of φ. Terms in the logic are interpreted as subsets of the 'universe' U, i.e. $[\![\varphi]\!] \subseteq U$. The interpretations of the operators and the ordering are given in equations (3.2) to (3.4).

$$[\![\varphi + \psi]\!] = [\![\varphi]\!] \cup [\![\psi]\!] \tag{3.2}$$

$$[\![\varphi \times \psi]\!] = [\![\varphi]\!] \cap [\![\psi]\!] \tag{3.3}$$

$$[\![\varphi \leq \psi]\!] = [\![\varphi]\!] \subseteq [\![\psi]\!] \tag{3.4}$$

This allows one to specify concepts in terms of other concepts, e.g. $ORANGE = YELLOW \times RED$ to say that orange is the intersection of yellow and red, or $GREEN = YELLOWISHGREEN \ + PROPERGREEN + BLUEISHGREEN$ to specify a bigger concept as the sum of its parts (both examples from (Fischer Nilsson 1999)). Furthermore, a null concept and a universal concept are defined as in (3.5) and (3.6), forming the bottom and the top of the lattice.

$$[\![\bot]\!] = \{\} \tag{3.5}$$

$$[\![\top]\!] = U \tag{3.6}$$

Obviously $\varphi + \bot = \varphi$ and $\psi \times \top = \psi$ hold.

To be able to reason about the quality dimensions of a conceptual space, Fischer Nilsson (1999) uses the Peirce product $(a:\varphi)$, which is the set of elements x related by (some relation) a to some element y in φ, defined formally in (3.7).

$$[\![a:\varphi]\!] = \{x \mid \exists y((x,y) \in a \wedge y \in [\![\varphi]\!]\} \tag{3.7}$$

Instead of $(a:\varphi)$ the notation $a(\varphi)$ is adopted, with a now being a unary operator that lifts a concept into a property. More formally, $a(\varphi)$ is the *inverse image* of the function that projects the concept φ onto the quality dimension a. This is illustrated in Figure 10, where first a concept φ is projected onto the a-axis as an interval φ', and then the inverse image $a(\varphi')$ is taken, which is a (hyper-)cylinder in dimensional space.

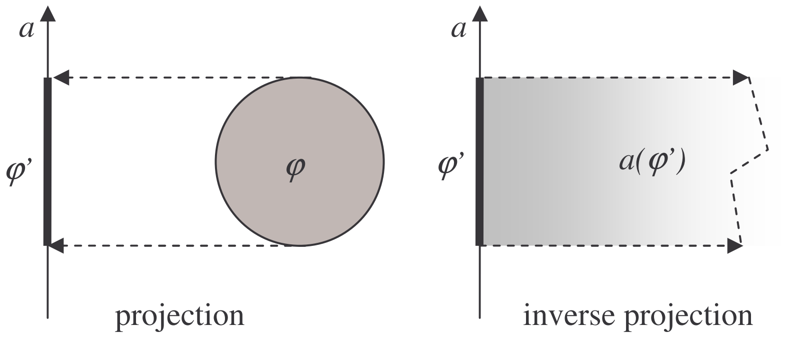


Figure 10. The interpretation of *a(φ)* as the inverse image of the projection of concepts onto *a*.

Using the Peirce product, a concept can be written as a *frame term*, where a base concept c is restricted by a list of properties $a_i(\varphi_i)$, as in (3.8).

$$c \times a_1(\varphi_1) \times a_2(\varphi_2) \times \ldots \times a_n(\varphi_n) \tag{3.8}$$

The base concept c may be the universal concept $\top$, in which case it can of course be omitted. The operators a_i correspond to the labels of the quality dimensions. For the list of properties I will adopt the shorthand notation in (3.9), as suggested in (Fischer Nilsson 1999).

$$c \times a_1(\varphi_1) \times \ldots \times a_n(\varphi_n) = c \times \begin{bmatrix} a_1 : \varphi_1 \\ \vdots \\ a_n : \varphi_n \end{bmatrix} \tag{3.9}$$

The Peirce product, crux operator and frame term construction are illustrated for a two-dimensional space in Figure 11.

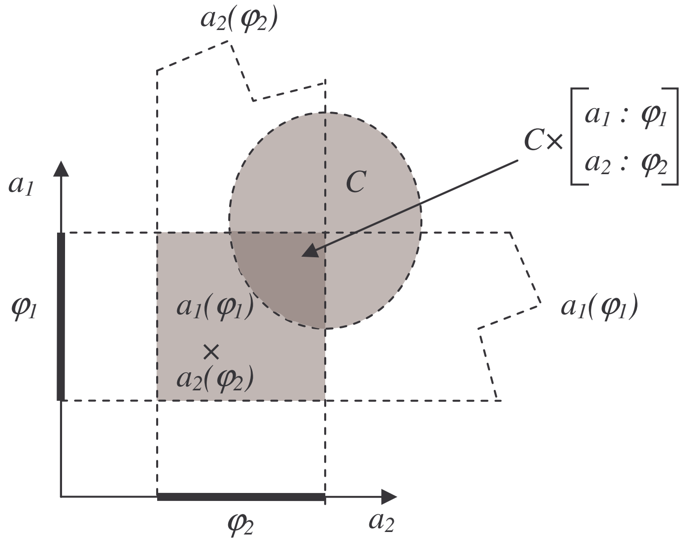


Figure 11. The construction of a frame term representing the part of concept C that has certain values for its attributes a_1 and a_2 (i.e. the dark quarter circle).

I will illustrate the notation introduced so far with some examples, adapted from (Fischer Nilsson 1999). The quality dimensions of the color spindle are *hue*, *saturation*, and *brightness*, the values of which I choose to be on the interval $[0,1]$. As an example, *RED* could correspond to a certain hue interval combined with any saturation and brightness. The definition of *RED* is given in (3.10). Of course the attributes that put no constraints on the concept, i.e. the attributes that are allowed to take any value in the domain, can be omitted in the description, since they constitute the universal concept $\top$.

$$RED = \begin{bmatrix} hue:[0.65,0.75] \\ saturation:[0,1] \\ brightness:[0,1] \end{bmatrix} = \begin{bmatrix} hue:[0.65,0.75] \end{bmatrix} \tag{3.10}$$

Definitions can also be given for concepts such as *BRIGHT*. The concept *BRIGHT* is defined as the collection of colors with a fairly high brightness, independent of the hue and saturation.

$$BRIGHT = \begin{bmatrix} hue:[0,1] \\ saturation:[0,1] \\ brightness:[0.5,1.0] \end{bmatrix} = \begin{bmatrix} brightness:[0.5,1.0] \end{bmatrix} \tag{3.11}$$

Similarly, we can define *DARK*, *STRONG* and *WEAK* as in equations (3.12) to (3.14).

$$DARK = \left[brightness:[0.0, 0.5]\right] \tag{3.12}$$

$$STRONG = \left[saturation:[0.5, 1.0]\right] \tag{3.13}$$

$$WEAK = \left[saturation:[0.0, 0.5]\right] \tag{3.14}$$

The concept *BRIGHTRED* can now be constructed as the part of *RED* that is also *BRIGHT*:

$$BRIGHTRED = RED \times BRIGHT = \begin{bmatrix} hue:[0.65, 0.75] \\ brightness:[0.5, 1.0] \end{bmatrix} \tag{3.15}$$

To account for a modular approach advocated in (Geuder and Weisgerber 2005), I will introduce some modifications that do not affect the logic of concepts within simple domains (e.g. the color domain) but which allow an extension to compound domains. Domains are regarded as tuples. Non-compound domains correspond to a singleton tuple containing only a reference to a primitive conceptual space, e.g. the color domain referring to the hue-saturation-luminance space: $color = \langle COLOR\_3D\_HSL \rangle$. Compound domains are tuples of domains, which may be simple or compound (i.e. domains can be combined recursively). For example, the higher level domain *attractiveness* might consist of the domains *appearance* and *personality*, the first of which is itself compound:

$$\begin{aligned} &attractiveness = \langle appearance, personality \rangle \\ &\quad = \langle \langle shape, texture, color \rangle, \langle CHARACTER9D \rangle \rangle \\ &\quad = \langle \langle \langle SHAPE5D \rangle, \langle TEXTURE3D \rangle, \langle COLOR3D\_HSL \rangle \rangle, \langle CHARACTER9D \rangle \rangle \end{aligned} \tag{3.16}$$

The 'points' in a hierarchical conceptual space are themselves hierarchical. A point in the *attractiveness* domain is in fact a tuple $\langle A, P \rangle$ where A is a 'point' in the *appearance* domain and P is a point in the *personality* domain. A itself is a tuple of points, one from each sub-domain: $\langle S, T, C \rangle$. S, T, C and P are then points in each conceptual space, of dimensionality 5, 3, 3 and 9 respectively. More formally, the interpretation of a concept in a compound domain is as follows:

$$[\![\varphi_A \wedge \ldots \wedge \psi_B]\!] = \{ \langle X, \ldots, Y \rangle \mid X \in [\![\varphi_A]\!], \ldots, Y \in [\![\psi_B]\!] \} \tag{3.17}$$

The interpretation of a concept in a non-compound domain remains a subset of the universe. But instead of speaking of the universe U of concepts, I will add a subscript to be able to speak of the universe U_{domain} referring to all the concepts of a certain domain. Analogously, concepts φ, including the universal concept $\top$, and also attributes a in the Peirce product notation, are given a domain subscript. To be able to reason with concepts across different domains, I introduce the domain function $Dom : D \times U_A \rightarrow U_D$ which takes a domain and a concept and returns the concept restricted to that domain:

$$[\![Dom(B, \varphi_A)]\!] = [\![\psi_B]\!] \text{ such that } \psi_B \text{ is } \varphi_A \text{ restricted to domain } B \qquad (3.18)$$

For example, the concept describing the color of apples can be written as the *APPLE* concept restricted to the color domain. I will adopt a shorthand notation as in (3.19), writing the domain argument itself as the functor.

$$Dom(color, APPLE_{fruit}) = color(APPLE_{fruit}) = \left(RED + YELLOW + GREEN \right)_{color} \qquad (3.19)$$

I adopt the logical conjunction symbol to construct concepts in composite domains. A sequence of such conjunctions can be seen as the syntactic counterpart of the tuple. A notation convention similar to the Peirce product in frame terms is adopted, such that complex concepts can also be written as a column vector which I call a *domain term*, as in (3.20).

$$\varphi_A = \bigwedge_{D \in A} D(\varphi_A) = D_1(\varphi_A) \wedge \ldots \wedge D_n(\varphi_A) = \begin{bmatrix} D_1 : D_1(\varphi_A) \\ \vdots \\ D_n : D_n(\varphi_A) \end{bmatrix} \qquad (3.20)$$

For example, if the fruit domain contains the domains color and shape (which is of course a simplification), *APPLE* can be described as in (3.21). Note that the concept subscripts are omitted in the domain term notation.

$$\begin{aligned} APPLE_{fruit} &= \bigwedge_{D \in fruit} D(APPLE_{fruit}) \\ &= \left(RED + YELLOW + GREEN \right)_{color} \wedge ROUND_{shape} \\ &= \begin{bmatrix} color : RED + YELLOW + GREEN \\ shape : ROUND \end{bmatrix} \end{aligned} \qquad (3.21)$$

The domain function returns the universal concept if the domain is not a part of the concept's complex domain, reflecting the fact that any value for that domain will do. For example, temperature is not a part of the shape property, so the temperature of the *ROUND* concept (which is totally nonsensical) is $temperature(ROUND_{shape}) = \top_{temperature}$.

With the modularity added to the logic, the terms become slightly more complex. I will adopt the strategy to first rewrite everything into constructions of non-compound domains, and then do the interpretation. In the following I will assume that set operations union and intersection apply to tuples and are order-preserving. If the domain $A \cup B$ is compound (i.e. $\left| A \cup B \right| > 1$) then we can do the following sum and crux rewriting (for non-compound domains the identities hold as well, but they will lead to circularity):

$$\varphi_A + \psi_B = \bigwedge_{D \in A \cup B} D(\varphi_A) + D(\psi_B) \tag{3.22}$$

$$\varphi_A \times \psi_B = \bigwedge_{D \in A \cup B} D(\varphi_A) \times D(\psi_B) \tag{3.23}$$

These identities are illustrated in Figure 12. Each row indicates a concept, the same domains being aligned vertically.

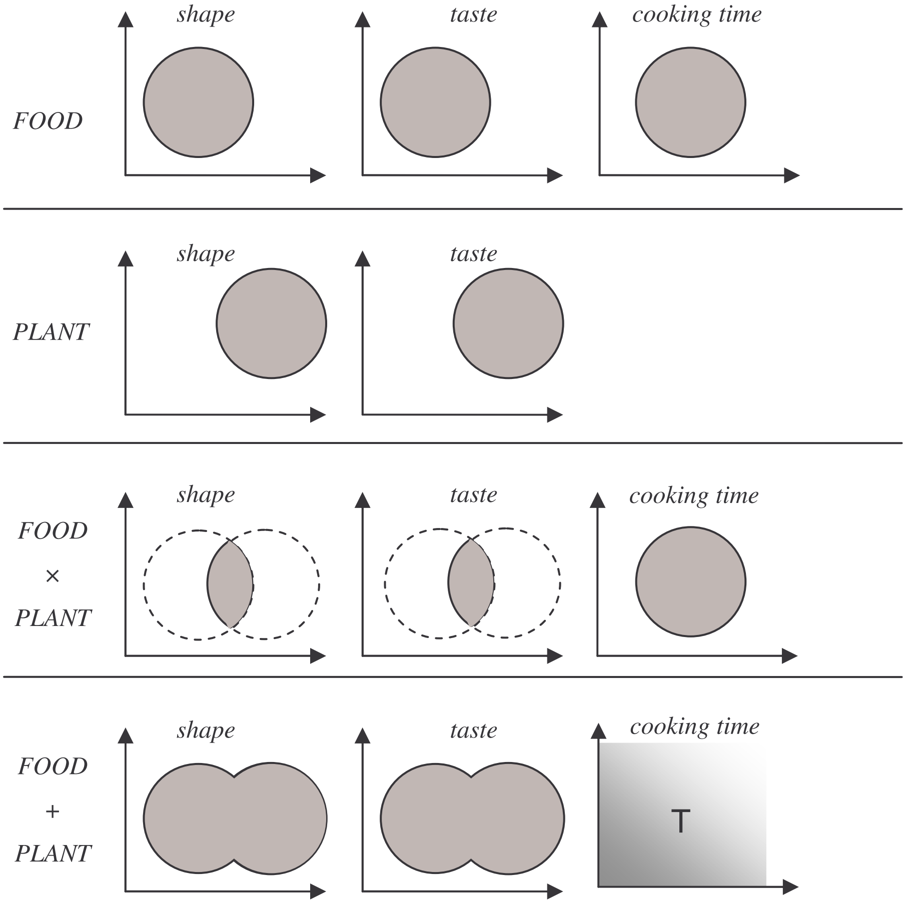


Figure 12. The crux and sum operations exemplified for compound domains.

To put this in words, the crux of two concepts is the concept of all things that belong to both concepts. The crux of *PLANT* and *FOOD*, two concepts that may live in different (possibly compound) conceptual spaces, refers to all the things that are both plant and food, i.e. all vegetables. The sum of two concepts is somewhat harder to grasp. The sum of *PLANT* and *FOOD* are all the possible plants and foods themselves, but it is much wider. As can be seen from the figure, it also includes all things that have the shape of food and the taste of a plant or vice versa. Of particular interest is the fact that in both operations, every subdomain is maintained. However, the sum operation results in a universal concept in the domains that are not shared (e.g. *cooking time*), meaning that $PLANT + FOOD$ can have any cooking time. Note also that with these rewriting rules in place it can be seen that $\varphi_A \wedge \psi_B = \varphi_A \times \psi_B$ and $\varphi_A + \psi_B = \top_{A \cup B}$ if A and B are disjoint (i.e. have no domains in common).

Note that domains that are nested at different depths in the domain hierarchy cannot 'see' each other in plus and crux operations. For example, combining a concept in the *appearance* domain with a concept in the *attractiveness* domain will not work, because none of the subdomains of *attractiveness* (*appearance* and *personality)* matches a subdomain of *appearance* (*shape*, *texture* and *color*). Instead, the concept from the *appearance* domain has to be 'lifted' to an equal level. Intuitively, this might correspond to the way humans combine concepts. We cannot sensibly intersect *RED* and *APPLE*, but we can intersect *RED THINGS* and *APPLE*. (Note that by this I do not wish to refer to the *linguistic* combination of concepts, e.g. in the constructs "red apple" or "fake banana", the meaning of which can be rarely captured by intersection).

The ordering relation needs some rewriting when occurring with compound concepts, this time quantifying over the component domains (again, it also works for singleton domains but it doesn't really *add* anything):

$$\varphi_A \leq \psi_B \Leftrightarrow \forall D \in \left(A \cup B\right).\left(D(\varphi_A) \leq D(\psi_B)\right) \tag{3.24}$$

To illustrate the ordering relation, let's see if $PLANT \leq FOOD$ holds. This holds only if for all domains that *PLANT* and *FOOD* together occupy, the part of *PLANT* restricted to that domain is a subconcept of the part of *FOOD* restricted to that domain. Obviously this can only hold if all the domains of *FOOD* are also domains of *PLANT*, which is not the case in the example due to the *cooking time* domain. Apart from that, suppose the two concepts have an *edibility* dimension in common. Obviously there are plants that are not edible and the ordering relation does not hold.

The interpretation of crux and sum remain the same, but they will only apply to singleton domains, that is if $\left|A \cup B\right| = 1$, so to interpret a concept one first needs to rewrite it through equations (3.22) and (3.23). The interpretation of the Peirce product remains fairly similar, as shown in equation (3.25). Note that, as before – and unlike the domain function – it returns a concept in the same domain as the original concept.

$$[\![a_A : \varphi_B]\!] = \{ x \mid \exists y((x, y) \in a_A \wedge y \in [\![A(\varphi_B)]\!] \} \tag{3.25}$$

So far the similarity and comparability measures have not been concretely defined. For now it is sufficient to say that the suggestion in (Geuder and Weisgerber 2005), given in equation (3.1), is a suitable approach. I will come back to similarity and comparability in section 3.4, after the required conceptual spaces have been formalized.

3.2 Motor representation

This section is threefold. The first part introduces a conceptual space for the representation of forces acting on joints, which are combined into a compound conceptual space. The second part discusses the expressivity of this approach and explains how the entities in effector space are to be combined into concepts representing complex motor patterns in a higher-level compound space which I will call *motor space*.

3.2.1 Joints and effectors

In chapter 2 I have shown that joints and effectors play key parts in the representation of motor patterns. All motion of effectors is grounded in the coordinated change of joint angles. For that reason these joints form the building blocks of the motor representation framework. A very basic type of motion of a joint is *roll*, the rotation of a joint along an axis. A roll of the joint is often accompanied by a *glide* inside the joint. As is depicted in Figure 13, a roll without a glide causes a displacement of the joint.

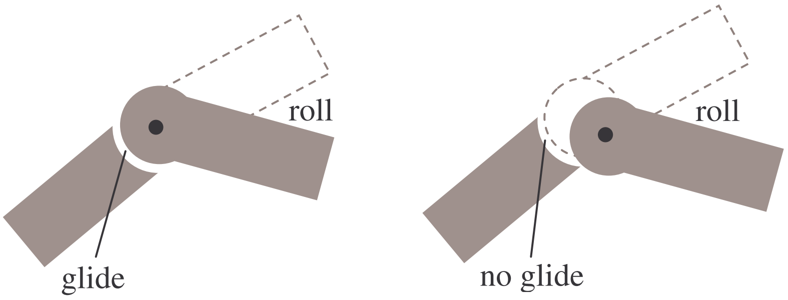


Figure 13. A schematic joint. Roll with glide (left) and without glide.

In general, the displacement of a joint is only very minimal, so in the model I will ignore this subtlety of the human body and assume that glide always occurs. Besides a roll and the corresponding glide, some joints allow the rotation of a joint around the axis of the connected bone, called *spin*. So a conceptual space is required that can represent the forces that cause these two types of joint movement: roll (with glide) and spin. See Figure 14 for a schematical depiction of a joint and the forces responsible. The joint is depicted in the origin.

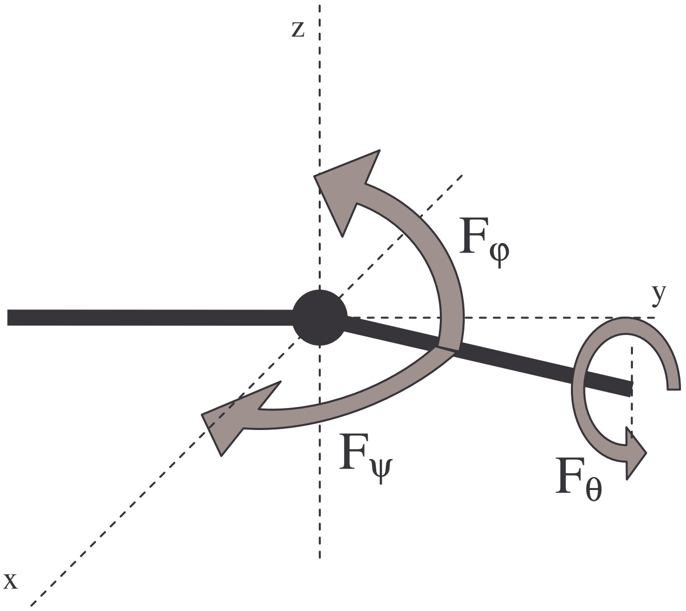


Figure 14. The two possible joint displacements and the three forces causing it.

To represent roll, the angular force F_φ in the vertical plane and the angular force F_ψ in the horizontal plane are sufficient. Spin can be described by just one parameter F_θ, which is simply the force about the axis. The generic vector describing the force at a joint thus resides in a three-dimensional space and is given by (3.26).

$$\varphi_{joint} = \begin{bmatrix} F_\varphi : \psi_1 \\ F_\psi : \psi_2 \\ F_\theta : \psi_3 \end{bmatrix} \tag{3.26}$$

This concept in *joint space* is very generic. It needs to be applied somehow to a concrete joint somewhere in the human body in order for it to be ‘complete’. To model this, I will regard concepts in joint spaces (and in section 3.3 analogously the concepts in property spaces) as functions that take as input an object and return a more specialized concept. I will simply write $\varphi_A(o)$ for a concept φ in domain A applied to an object o. I will show later how this functionality is propagated through composition of domains and their concepts.

As an example, the flexion force on a joint in the vertical direction (i.e. upwards roll), for example in the shoulder when raising an arm, can be represented by the universal joint concept restricted to those with positive values for F_φ, written as a frame term:

$$POSROLL_{joint} = \left[F_\varphi : \mathbb{R}_+ \right] \tag{3.27}$$

Of course I had to decide first on whether the positive value corresponds in the case of a shoulder to an up or down force (in this case up). A spin force around a joint, in a clockwise direction, which is also, to a certain degree, anatomically possible for joints like the shoulder, can be described as follows:

$$POSSPIN_{joint} = [F_\theta : \mathbb{R}_+] \tag{3.28}$$

These concepts are still very generic. We can apply them to the shoulder to give them real meaning. We can also combine these two concepts to denote a more complex force composition, i.e. one that when applied to the shoulder joint would result in a strange force that might be useful only in a very primitive type of tennis (I encourage the reader to try):

$$SLAM_{joint} = POSROLL_{joint} \times POSSPIN_{joint} \tag{3.29}$$

Note that these forces, for example applied to the shoulder, do not guarantee that the attached arm will actually slam – maybe there are other forces that hold it back or even move it in the opposite direction. This is not the resultant force.

All physical human actions are somehow performed by effectors, the parts of the human body that are used for interaction with the world. Parts that are often used as an effector are the hands and feet, but for some actions the elbow, the head or even a shoulder may be considered the effector. The forces applied by effectors are thus of key importance for the performance of an action. The forces applied by an effector are constituted by a set of joints (the effector system). It is plausible that the joints-effector hierarchy is reflected in our conceptualization of motor patterns. I propose a compound conceptual space of the type argued for in (Geuder and Weisgerber 2005), which I will call *effector space*, to represent the forces applied by effectors. In modular conceptual space logic, the domain would be $effector = \langle joint, joint, ... \rangle$. This hierarchy is illustrated in Figure 15.

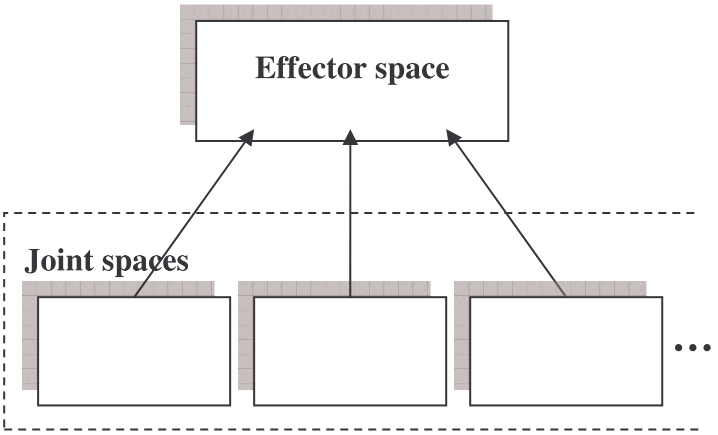


Figure 15. The domains of the compound conceptual effector space for the human hand.

Note that concepts in this effector space do not explicitly represent the force applied by the effector. However, it follows from the unique decomposition of the concepts into joint space concepts that every concept in effector space still represents a particular (set of) forces. Of course, the same effector force may be represented by different concepts in effector space, because a different configuration of joints may lead to the same result. Try pressing the palm of your hand against the tabletop, and then the back of your hand. The same force can be applied in two ways with different joint configurations. This joint redundancy is sometimes seen mainly as a challenge for which humans have somehow developed effective heuristics. However, from a robotic viewpoint, this redundancy has also been viewed as a property increasing flexibility and adaptability, for example in (Kreutz-Delgado, *et al.* 1992).

Even though this "effector space" is a compound space, I will refer to it as a regular Gärdenforsian conceptual space, with points (that are actually tuples of points) and regions (which are actually tuples of regions in the component spaces). Just to avoid confusion with the *knoxels* in the approach of (Chella, *et al.* 2000), I will (somewhat awkwardly I admit) call the 'points' in effector space *foxels*, because they represent forces.

As I explained, concepts in joint space are generic concepts, functions yet to be applied. How does this property transfer to concepts in the higher, compound effector space? Concepts in effector space, or sets of foxels, are functions that need to be applied to an effector, such as the hand or the foot. The joint space concepts of which it consists are then applied to the joints in the corresponding effector system. For example, take a generic concept $RAISE_{effector}$ in effector space, consisting of a $POSROLL_{joint}$ (bending) and a $NEGROLL_{joint}$ (stretching) in different joint spaces:

$$RAISE_{effector} = POSROLL_{joint} \wedge NEGROLL_{joint} \tag{3.30}$$

When we apply this generic motor representation to the hand effector, the application propagates further down the hierarchy to the joints that are responsible for hand movement:

$$RAISE_{effector}(hand) = POSROLL_{joint}(shoulder) \wedge NEGROLL_{joint}(elbow) \tag{3.31}$$

Or when applied to the foot effector:

$$RAISE_{effector}(foot) = POSROLL_{joint}(hip) \wedge NEGROLL_{joint}(knee) \tag{3.32}$$

This kind of downward propagation of concept application requires knowledge of which joints make up which effector system. Also, it must be known which joint applies to which joint space concept, maybe through some kind of a hierarchical ordering of joints. For applications in artificial intelligence, obviously this knowledge has to be specified. But I do not foresee any difficulties in this and will not treat it here. For an example, see the framework for imitation learning in (Chella, *et al*. 2006).

It is obvious how the effector-independency of motor representations, discussed in chapter 2, is incorporated in this representation. Concepts in effector space may be applied to any effector. As equation (3.32) displays, the concept of raising a hand (shoulder, elbow) can be readily transferred to raising a foot (hip, knee). However, not all effector systems are compatible. When applying concepts that are constructed foremost for the hand effector system to, say, the elbow effector system, the transfer cannot be very good because there is simply one joint lacking in the latter system. Effector-independency does not guarantee compatibility, neither in this model nor in our brains. But full compatibility is not required for transfer, as was illustrated by the five writings of "cognitive neuroscience" in Figure 5 in the previous chapter. The motor plan for writing was reasonably successfully transferred from the hand to the mouth, an effector system with a wholly different structure.

Now, as a slightly more complex example of a concept in effector space, consider the action of serving (as in tennis). I encourage the reader to try the different parts of the motion the way I did while writing. To prepare the serve, you raise your arm while slightly twisting it, while bending your elbow (this is of course approximate). Let's start with the concepts for flexion (positive F_{φ} value) and counterclockwise spin (negative F_{θ} value) of the shoulder. For simplicity I will only use the values 'positive' or 'negative', though in general only a narrow interval of force magnitudes will be suited for an action. The concepts are denoted by the following frame terms:

$$POSROLL_{joint} = \left[F_{\varphi} : \mathbb{R}_{+} \right] \tag{3.33}$$

$$NEGSPIN_{joint} = \left[F_{\theta} : \mathbb{R}_{-} \right] \tag{3.34}$$

This is combined into the concept denoting a simultaneous flexion and negative spin force of a joint, which, of course, still resides in the $joint$ space:

$$RAISE_{joint} = POSROLL_{joint} \times NEGSPIN_{joint} \tag{3.35}$$

Simultaneously the elbow has to bend, which is in this case the same concept as *POSROLL* but applied to the elbow. In reality, such a re-use of concepts is probably rarely possible, but here I work only with a very coarse approximation of the 'true' concepts.

Now the concept in effector space (i.e. a set of foxels) *RAISE&BEND* can be constructed:

$$RAISE\ \&\ BEND_{effector} = \begin{bmatrix} joint : RAISE \\ joint : BEND \end{bmatrix} \tag{3.36}$$

And when we apply this generic motor concept to the right hand effector, it looks like this:

$$RAISE\ \&\ BEND_{effector}(righthand) = \begin{bmatrix} joint : RAISE(shoulder) \\ joint : BEND(elbow) \end{bmatrix} \tag{3.37}$$

Which represents the raising, twisting and bending of the right hand effector system during the first phase of the serve.

It is time to mark an important difference to the approach of (Chella, *et al.* 2000). The knoxels in their conceptual space do not represent the same entities as foxels in the space described here. Knoxels represent 'simple motions' (i.e. position as a function of time), whereas the foxels in my approach represent something like a uniform *acceleration* (i.e. velocity as a function of time), at least with respect to the joints. In principle, the conceptual space used in (Chella et al. 2000) is more expressive than the one proposed here. They use the *Fourier transform* to be able to represent the position of an object as any arbitrary function of time. My force-based approach, on the other hand, only allows position functions that depend on time *quadratically*. Fortunately, for human bodily motion this is no real restriction. Most human motion patterns consist of uniformly accelerating and decelerating phases, generated by a constant force – and such phases are perfectly described by (maximally) quadratic functions. Higher-order polynomial functions seem only required for skilled dancers. Note also that even though the forces that work on the joints are all constant, the force on the effector resulting from it need not be constant.

To represent the *simultaneous* motor patterns of several effector systems, e.g. throwing a ball up in the air with your left hand while raising a racket with your right hand, another level is added to the hierarchy. This level, depicted in Figure 16, is composed of effector spaces and is called *segment space*, a naming that will become clear in just a few pages. Because the same composite force for different durations has completely different effects, the duration of it is included in the representation (which is just a one-dimensional vector with a real number). The choice to include the duration at this level (why not at the joint level?) will also make more sense in a page or two. The compound “points” in segment space, which may be adequately called *composite foxels*, represent the simultaneous motor patterns of the component effectors. In this hierarchical approach, it is hard to see that the compound spaces can still be regarded as conceptual spaces, provided that a distance function is defined to deal with the modularity. As I announced before, this will be treated in section 3.4.

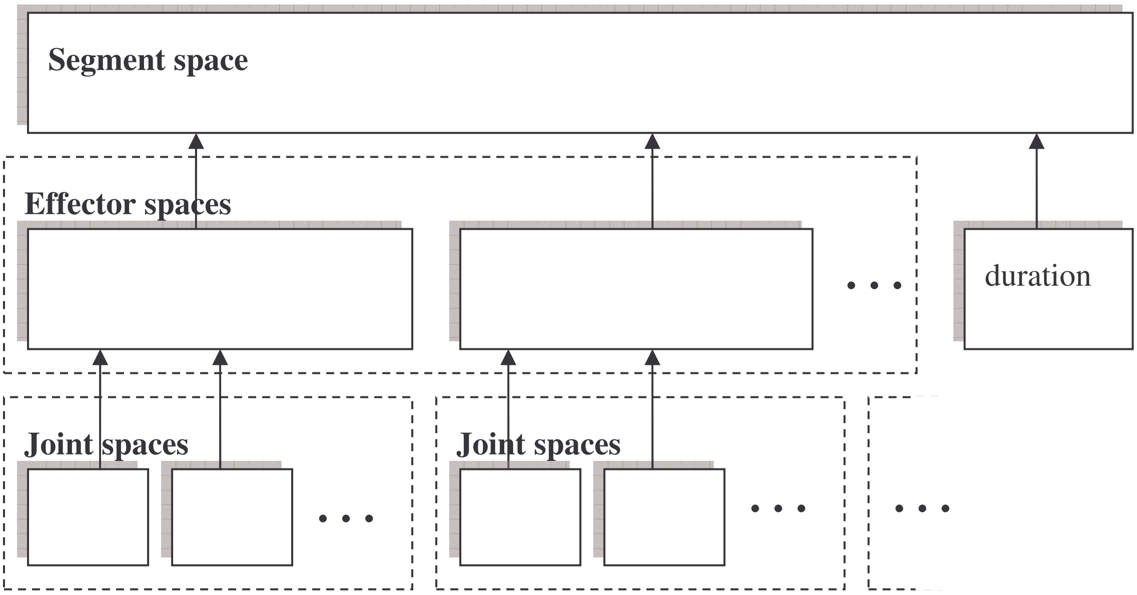


Figure 16. Segment space is composed of effector spaces representing motor patterns of different effector systems, plus a duration space.

Consider once again the tennis serve. As an illustration I will construct the motor representation of the left hand (throwing the shuttle in the air), that occurs simultaneously with the motor pattern of the right hand. For simplicity I consider it a simple raise of the arm through a positive roll applied to the shoulder (equation (3.38)).

$$LIFT_{effector} = \left[joint : POSROLL \right] \tag{3.38}$$

The left hand and right hand concepts can now be combined together with a duration concept, (let's assume the concept *SHORT* exists) into a concept in segment space, which denotes the entire first phase of the serve (ignoring leg movement, for now):

$$SERVE\_PREPARE_{segment} = \begin{bmatrix} effector : RAISE \;\&\; BEND \\ effector : LIFT \\ duration : SHORT \end{bmatrix} \tag{3.39}$$

This concept contains all the information, down to the lowest level, about the (constant) joint forces in two effector systems. Still, it is very generic. It may be applied to a man or a woman, or even to a monkey or a duck (with only little compatibility). Again, function application propagates downwards:

$$SERVE\_PREPARE_{segment}(human) = \begin{bmatrix} effector : RAISE \;\&\; BEND(righthand) \\ effector : LIFT(lefthand) \\ duration : SHORT \end{bmatrix} \tag{3.40}$$

Of course, when the concept is applied to a left-handed human, function application propagates slightly differently (and it will look slightly odd to right-handed people). This concept represents the full motoric part of the first segment of the tennis serve, as performed by a right-handed human.

3.2.2 *Motor space*

As I have shown in chapter 2, the use of forces instead of velocities makes the elegant state-motion-state decomposition presented in (Marr and Vaina 1982) and adopted in (Chella, *et al.* 2000) somewhat problematic. Instead of motion itself, the accelerations/decelerations can be segmented. When acceleration (of joint motion) changes in direction or magnitude, the point in the joint space representing it is displaced, shifted to a different position in the joint space. Most actions seem to consist of a periodical alternation of acceleration/deceleration, and this then corresponds to a periodical shifting of the points in joint spaces. (Again attention must be paid: this shifting is similar to the 'scattering of knoxels' in (Chella, *et al.* 2000), but it represents far from the same process.) With this in mind, it makes sense to add the duration representation to segment space rather than to any other level of representation: it is the level at which *motor segments* are represented.

So a complex motor pattern is in fact a series of segments of constant (and possibly zero) acceleration/deceleration for the component effectors' joints. Each segment is represented by a composite foxel in the conceptual space that is composed of a set of effector spaces, each representing a different effector. It is of course possible that in one segment the left hand and right foot are involved whereas in another segment the head and the left foot are. Segments are separated from one another by a shift of the composite foxel (i.e. a shift of one or all of the points in the joint spaces). To represent this hierarchy I add yet another, higher-level conceptual space that is compound, composed of the several segment spaces describing different segments and their durations, still sticking to the modular approach advocated in (Geuder and Weisgerber 2005). This compound conceptual space, which I call *motor space* because it represents the most complete motor patterns (there will be no higher motor representation level after this), is depicted in Figure 17.

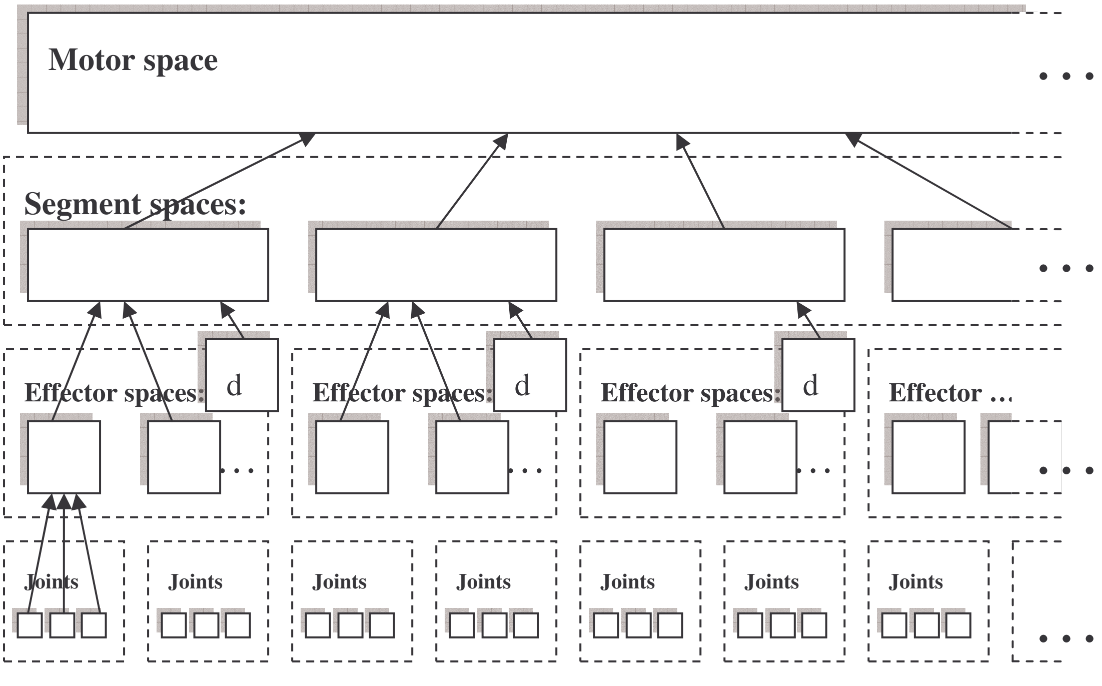


Figure 17. At the highest level, the compound motor space is composed of the segment spaces that represent all the effector systems of the human body.

The compound points in motor space represent an arbitrary number of motor segments. As a final exercise I will construct a concept representing the entire motor pattern of the tennis serve. Remind that we already constructed the concept for the first segment in (3.39), repeated here:

$$SERVE\_PREPARE_{segment} = \begin{bmatrix} effector: RAISE \,\&\, BEND \\ effector: LIFT \\ duration: SHORT \end{bmatrix} \tag{3.41}$$

The second segment (which is in fact *VERYSHORT*) consists of a drop of the left arm, to balance the body, and a slam with the right arm, represented by concepts in effector space, as follows:

$$DROP_{effector} = \left[joint: [\psi : \mathbb{R}_{-}] \right] \tag{3.42}$$

$$SLAM_{effector} = \begin{bmatrix} joint: \left[F_{\varphi} : \mathbb{R}_{-} \right] \\ joint: \left[F_{\varphi} : \mathbb{R}_{+} \right] \end{bmatrix} \tag{3.43}$$

$$SERVE\_HIT_{segment} = \begin{bmatrix} effector: SLAM \\ effector: DROP \\ duration: VERYSHORT \end{bmatrix} \tag{3.44}$$

The whole action of two-armed serving is represented by the following concept in motor space, applied to a human:

$$SERVE_{motor}(human) = \begin{bmatrix} segment: SERVE\_PREPARE(human) \\ segment: SERVE\_HIT(human) \end{bmatrix} \tag{3.45}$$

Note that in the current framework there is no way to formally describe the sequential order of the component spaces. A *sequence* operation could be added to the conceptual space logic that takes two concepts and is interpreted as a set of ordered pairs, or the segments could be numbered – but these might be nothing but ad-hoc solutions. A solution has to be found before the model can be implemented, but in this thesis I will skip such a formality.

3.3 Goal representation

In this section I will formalize a composite conceptual space in which the goal of an action can be represented. This section is divided into three parts. The first part discusses how various relevant properties of objects may be represented in a composite conceptual space called *state space*. The second part formalizes this. The third part shows how goals may be represented, both as an endstate and as a state change, in the higher-level composite goal space.

3.3.1 Properties

As I have mentioned in chapter 2, objects have many different properties that may change as a result of an action. An iron bar handled by a smith can change state with respect to shape, position, color, temperature and various other properties. Each of these properties can be represented in a pretty straightforward conceptual space. Temperature can be represented on a one-dimensional scale from zero Kelvin to infinity. Color can be represented in the three-dimensional color spindle. Position can be represented by the three Cartesian coordinates x, y, and z with respect to a chosen origin. Orientation is fully specified by three orientation parameters α, β and γ. Shape can be represented roughly by specifying the length, width and height. See Figure 18 for some examples of what I will call *property spaces*.

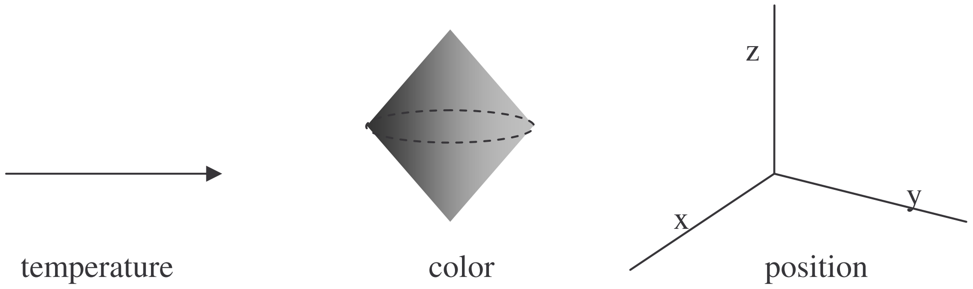


Figure 18. Some examples of property spaces.

The quality dimensions for most of these properties are straightforward. However, the quality dimensions for representing shapes (in a more detailed way than just the length, width and height) are less trivial. I here discern between three sorts of shape: global shape, material structure and functional structure. The first concerns global descriptions of length, width, and curvature. The second concerns the type of material, e.g. wood or glass. The third concerns the functional components, e.g. the ear of a cup and the handle of a hammer, and their relative orientations and positions within an object.

The third aspect, decomposition into functional components, seems to be the most expressive. Various approaches exist. In (Marr and Nishihara 1978) the problem of shape representation is tackled using cylinder models like the ones used for motion representation in (Marr and Vaina 1982). An important aspect of this approach is the assignment of a coordinate system to each cylinder and the recognition of the natural axes of shapes. A similar but more expressive approach is presented in (Biederman 1987), called *recognition-by-components*, which is based on the decomposition of objects not into cylinders but into elementary cone-like shapes or *geons*. Geons can be extracted from a two-dimensional image by discrimination on five detectible properties of edges, such as collinearity and parallelism. The model also makes predictions regarding the human faculty for object recognition, which are empirically validated.

Complex shapes may be represented in various ways, but it is, in the case of goals, the *change* of these shapes that I am interested in. Even the most complex shapes can change, as long as they are rigid, only in a small number of ways (e.g. flatten, lengthen, bend, break), so it seems that a shape representation need not concern functional decomposition, an expressivity only required for actions involving complex instruments or patients with moving parts. I foresee no problems in making such a property space explicit, but it is not required for the aims of this thesis. Instead, the focus here will be on the remaining two aspects: global shape and material structure. To represent the first, shapes will be approximated by elementary cubic objects, or *superquadrics*. Informally, superquadrics are all possible three-dimensional shapes that occur when transforming a box into a sphere, e.g. from a box, to a rounded box, to a sphere. I will present a more formal account later. To represent the second shape aspect, the material structure of objects, a wholly different property space is required.

Figure 19 shows an example of a property space to represent material structure. The property space in the figure is composed of two somewhat vague dimensions, which I shall explain. Let's say that both dimensions range conveniently between zero and one. Vertical in the figure is the *wholeness* dimension. A one value on that axis denotes that the object exists as a whole in reality. A zero value denotes that the object in fact does not exist (i.e. is completely disintegrated). It seems odd to speak of objects that do not exist, but in fact we do it all the time when we describe a *change* in the structure of an object. This change can be from existence to non-existence, e.g. "the plank broke into two parts" and "the glass scattered", but

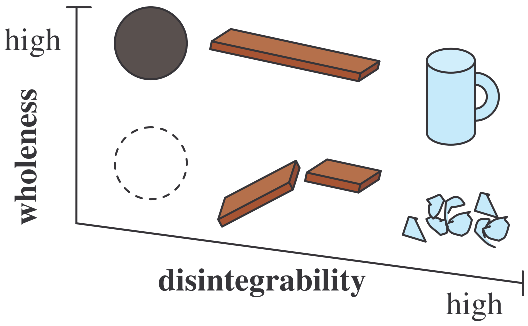


Figure 19. A two-dimensional property space for material structure.

also in the other direction, e.g. “the plank was created from two smaller ones” and “glass is synthesized by heating the cullet”.

The horizontal dimension in the figure is called *disintegrability* and represents the way in which an object disintegrates when it changes from high to low wholeness. A wooden plank generally disintegrates into two parts (unless if you use a shredder). A metal ball doesn’t disintegrate at all, so a change in it’s wholeness would have to amount to the disappearance of the ball. This provides us with a way to represent the meaning of verbs such as “disappear” and “appear” as paths in this space. A glass plate on the other hand disintegrates into many (but still countable) parts, a fact which becomes apparent when you hit it and it shatters, so it has a value somewhere in the right half of the disintegration scale. Water can be said to disintegrate into practically infinitely many components (waterdrops), represented by a disintegrability of one. Of course everything will scatter if you hit it with enough force - but let us stay in the domain of the probable, the domain in which our brains took shape.

The goal of an action can be represented in property spaces. Which property space should be addressed depends on the type of the action. For pushing and pulling, the position space seems suitable. For heating and cooling the property space of temperature is all that is required. For spatial tasks, a combination of position, orientation and possibly shape space seems the right candidate for goal representations. Arbitrarily many properties may of course be combined to represent goals. To account for this, the modular approach advocated by Geuder and Weisgerber (2005) and discussed in section 3.1 above comes in handy again. The modules or domains are in this case the different property spaces.

Which properties, i.e. which domains, are considered relevant in this study? The model should at least be able to represent position, orientation and shape (global shape and material structure), for these are the most basic (and most informative) properties of objects in general. In addition, to be forthcoming to the smith that appears so often in our examples, I will include the property temperature to enable him to heat the iron. This renders a compound conceptual space as in Figure 20.

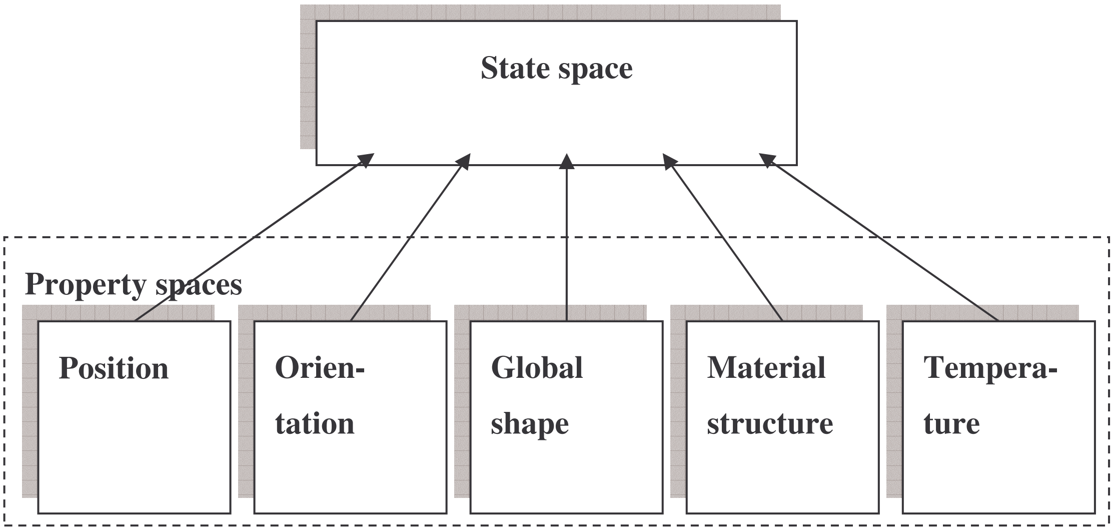


Figure 20. The domains of the compound conceptual state space.

Like the conceptual space in the previous section, I will refer to this compound space as a regular Gärdenforsian conceptual space, which I will call *state space*. A ‘point’ in state space, which is in fact a tuple of points in several property spaces, represents the state of an object with respect to the component properties. In the previous section I introduced the term foxels for points in effector space. Accordingly, the points in state space will be called *stoxels*.

As the focus in this thesis is on purely physical actions involving no sentient patients, mental states need not be considered (e.g. a change in happiness as a result of hitting). However, it is certainly possible to extend this framework to include psychosocial forces as well as mental state changes, assuming the appropriate quality dimensions can be found (e.g. “angriness” and “happiness”). Such flexibility is a major advantage of this modular approach.

3.3.2 Formalization of state space

So the state s of an object must be representable in five domains: position (pos), orientation (or), global shape (gs), material structure (ms) and temperature (tmp). The generic stoxel can be described with the domain term in (3.46), but often not all domains will be part of the state representation.

$$\varphi_{state} = \begin{bmatrix} pos:\psi_1 \\ or:\psi_2 \\ gs:\psi_3 \\ ms:\psi_4 \\ tmp:\psi_5 \end{bmatrix} \tag{3.46}$$

As in section 3.2 above I will consider such generic concepts as functions that require an object in order to be ‘complete’.

Position is represented by three parameters indicating the position of the object with respect to the agent, in Cartesian coordinates $x, y, z \in \mathbb{R}$. For the sake of clarity I will consider the agent’s center aligned with the origin, with its ‘front’ pointing along the x-axis and its top side along the z-axis, such that the unit vector $\begin{bmatrix}1 & 1 & 1\end{bmatrix}^T$ corresponds to a location in front, to the right and above the agent. The generic frame term is given in (3.47).

$$\varphi_{pos} = \begin{bmatrix} x:\psi_1 \\ y:\psi_2 \\ z:\psi_3 \end{bmatrix} \tag{3.47}$$

For orientation I will use the *Euler angles* approach, in which the orientation of a coordinate system (X, Y, Z) (of an object) is described with respect to a fixed system (x, y, z), in our case the system of the agent, as in Figure 21. The *line of nodes* is defined as the intersection of the xy-plane and the XY-plane. The parameter α describes the angle between the x-axis and the line of nodes. A second parameter, β, is the angle between the z-axis and the Z-axis. α and γ can have all values on the interval $[0, 2\pi]$. β can have all values on $[0, \pi]$.

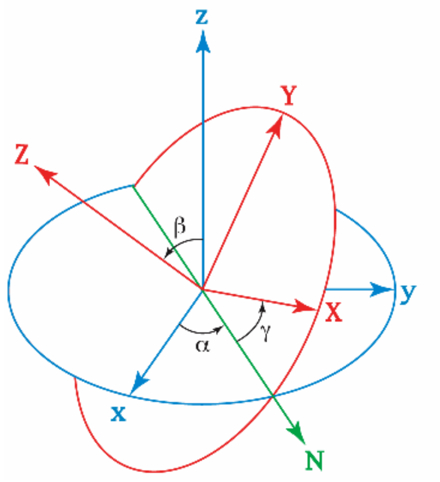


Figure 21. An illustration of the Euler angles. From Wikipedia, author: L. Brits (2008).

To determine the orientation of an object, as well as the global shape, one needs a way to actually determine the axes of an object. This is not trivial, because some objects seem to have an inherent coordinate system that other objects completely lack. In principle, rules of thumb could be applied, e.g. “the X-axis of a hand tool in use always points in the direction of the lower arm”. This is a general issue in robotics and machine vision which I will not go into here. The generic orientation frame term is, quite trivially, as given in (3.48).

$$\varphi_{or} = \begin{bmatrix} \alpha : \psi_1 \\ \beta : \psi_2 \\ \gamma : \psi_3 \end{bmatrix} \tag{3.48}$$

The next property is the global shape of the object. I argued above for an approach based on the superquadric approximation of shapes. A superquadric can be described in a parametric form, as a vector in three-dimensional space, as follows:

$$f(\eta, \omega) = \begin{bmatrix} a_x \cos^{\varepsilon_1} \eta \cos^{\varepsilon_2} \omega \\ a_y \cos^{\varepsilon_1} \eta \sin^{\varepsilon_2} \omega \\ a_z \sin^{\varepsilon_1} \eta \end{bmatrix} \tag{3.49}$$

Where $-\pi/2 \leq \eta \leq \pi/2$ and $-\pi \leq \omega \leq \pi$. The parameters describe the lengths of the three axes $a_x, a_y, a_z \in \mathbb{R}^+$ and the form $\varepsilon_1, \varepsilon_2 \in [0,1]$. Figure 22 illustrates in what way the form parameters influence the shape of a superquadric. Informally, the form parameters $\varepsilon_1, \varepsilon_2$ describe the ‘roundness’ of the object horizontally and vertically.

To represent the curvature of a superquadric, three parameters c_x, c_y, c_z are added that describe the curvature of the x-axis in the xy-plane, of the y-axis in the yz-plane (see Figure 23) and of the z-axis in the zx-plane. The curvature parameters describe the curvature of an axis in radians per length unit, and may take any real value. For example, a curvature of

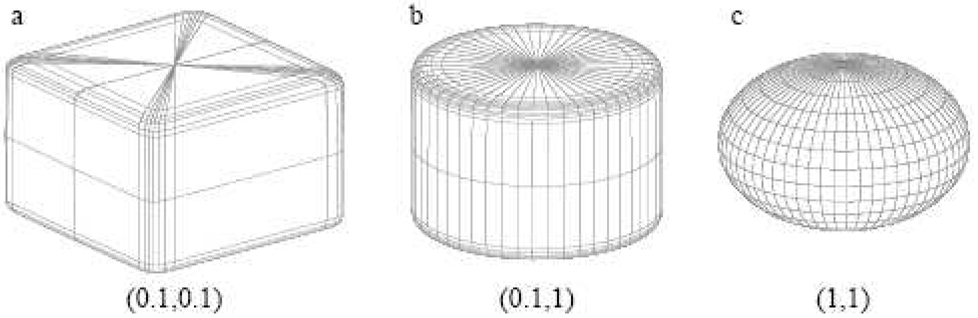


Figure 22. Different shapes of superquadrics when varying the form factors. Taken from (Chella, *et al.* 2000).

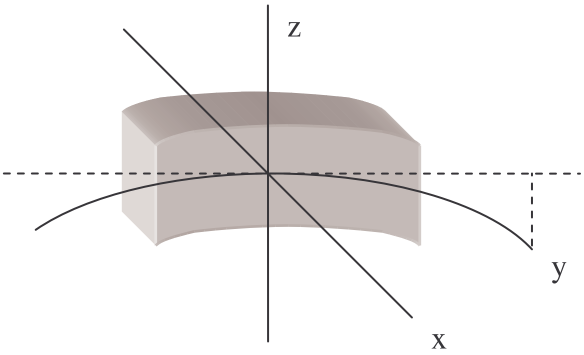


Figure 23. An example of a curved superquadric with $c_y < 0$.

$-\pi$ means that a superquadric of length 1 describes half a circle (π radians) in the negative direction of the axis. In summary, the generic frame term representing global shape is given in (3.50). For display purposes the frame term is shown as the transpose of a row vector, indicated by the superscript T.

$$\varphi_{gs} = \begin{bmatrix} a_x : \psi_1 & a_y : \psi_2 & a_z : \psi_3 & \varepsilon_1 : \psi_4 & \varepsilon_2 : \psi_5 & c_x : \psi_6 & c_y : \psi_7 & c_z : \psi_8 \end{bmatrix}^{\mathsf{T}} \tag{3.50}$$

The material structure of the object is described by the two parameters I introduced above, with $w \in \mathbb{R}_+$ describing the wholeness of the object and $d \in \mathbb{R}_+$ the disintegrability:

$$\varphi_{ms} = \begin{bmatrix} w : \psi_1 \\ d : \psi_2 \end{bmatrix} \tag{3.51}$$

The final property, temperature, is simply represented by a positive real number $t \in [0, \rightarrow\rangle$, i.e. as the following one-dimensional vector:

$$\varphi_{tmp} = [t : \psi] \tag{3.52}$$

With the formal definition of each property in place, concepts may be defined by putting constraints on the attribute values of a domain using the frame term notation, e.g. *INFRONT* (a positive x-coordinate) (3.53) and *UPRIGHT* (3.54) (the angle between the z-axis and the Z-axis is near zero). Whenever no confusion can occur (e.g. it is clear that the attributes x, y, z refer to the dimensions of the position space), subscripts will be omitted.

$$INFRONT_{pos} = \begin{bmatrix} x : \langle 0, \rightarrow\rangle \\ y : \mathbb{R} \\ z : \mathbb{R} \end{bmatrix} = [x : \langle 0, \rightarrow\rangle] \tag{3.53}$$

$$UPRIGHT_{or} = \begin{bmatrix} \alpha : [0, 2\pi] \\ \beta : [0, 0.1\pi] \cup [0.9\pi, \pi] \\ \gamma : [0, 2\pi] \end{bmatrix} = \left[\beta : [0, 0.1\pi] \cup [0.9\pi, \pi]\right] \tag{3.54}$$

These concepts can then be used in the description of states. For example, one might say that the intended endstate of walking is to be on a position in front of the current one whilst staying upright (since falling forward is not considered a successful realization of walking). The state of the body that is meant to result from walking can be represented by a concept in the state space, i.e. a tuple of stoxels, that is constructed as follows:

$$\begin{aligned} WALKGOAL_{state} &= INFRONT_{pos} \times UPRIGHT_{or} \\ &= \bigwedge_{D \in pos \cup or} D(INFRONT_{pos}) \times D(UPRIGHT_{or}) \\ &= \left[x_{pos} : \langle 0, \rightarrow \rangle\right] \times \mathrm{T}_{pos} \wedge \mathrm{T}_{or} \times \left[\beta_{or} : [0, 0.1\pi] \cup [0.9\pi, \pi]\right] \\ &= \left[x_{pos} : \langle 0, \rightarrow \rangle\right] \wedge \left[\beta_{or} : [0, 0.1\pi] \cup [0.9\pi, \pi]\right] \end{aligned} \tag{3.55}$$

Or, written as a domain term with frame terms as values:

$$WALKGOAL_{state} = \begin{bmatrix} pos : \left[x : \langle 0, \rightarrow \rangle\right] \\ or : \left[\beta : [0, 0.1\pi] \cup [0.9\pi, \pi]\right] \end{bmatrix} \tag{3.56}$$

After some rewriting this turns out to have exactly the intended meaning. The interpretation is the set of pairs of points, one point in the positive-x region of the position space and one point in the upward-β region of the orientation space.

3.3.3 Goal space

I will now describe in more detail how state space may be used to describe the goal of an action. Consider again the independent control model from chapter 2, in Figure 24. This time the spatial terms are replaced by the more generic terms “endstate” and “state change”.

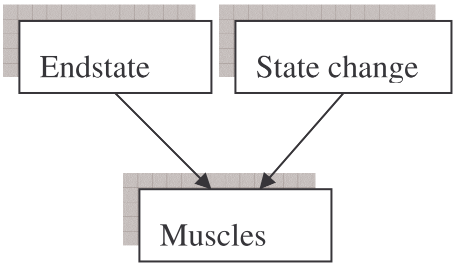


Figure 24. The independent control model for motor planning.

To be able to represent the goal both as an endstate and as a state change, a small change is required in the way state space has been viewed so far. The state space discussed represents the state of the object. It can be easily made into a *state change space* by considering the points in the space to represent state changes instead of states. Accordingly, I will discern between property spaces and property change spaces that form the base of the higher-level spaces. So, whereas a point in position space somewhere on the positive x-axis denotes a position in front of the agent, a point in position *change* space on the positive x-axis denotes a position *change* in the forward direction, regardless of the object's actual position with respect to the agent. I will denote change domains with an arrow. Compare $INFRONT_{pos} = [x : \mathbb{R}_+]$, which denotes all positions in front of the agent, with $INFRONT_{\overrightarrow{pos}} = [x : \mathbb{R}_+]$, which denotes all forward changes in position of the object. This crucial difference is illustrated in Figure 25.

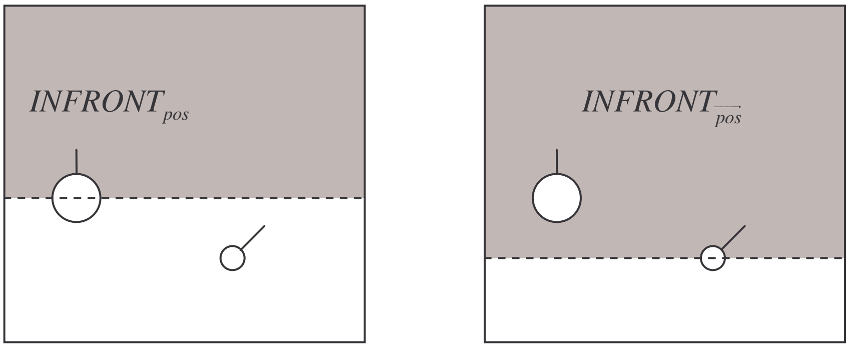


Figure 25. The difference between a concept in position space and a concept in position change space (right).

This kind of distinction might not capture the true nature of the independent control model, which concerns the distinction between an endpoint and a trajectory. Points in position change space do not really specify trajectories, they only represent *distances*. This is a simplification that is convenient for the types of actions considered in this thesis. It is only when adpositions are added to the picture that trajectories other than those fully specified by distances start to play a role. In that case, the notion of a path through state space would have to be specified and added to the framework. I will omit such an extension here.

For various actions, the goal may concern the states of several objects (i.e. instruments and patients). Therefore, instead of representing the goal in state (change) space itself, I define a higher-level compound conceptual space, *goal space*, that supervenes on the state spaces of all objects involved in an action. This modular structure is depicted in Figure 26.

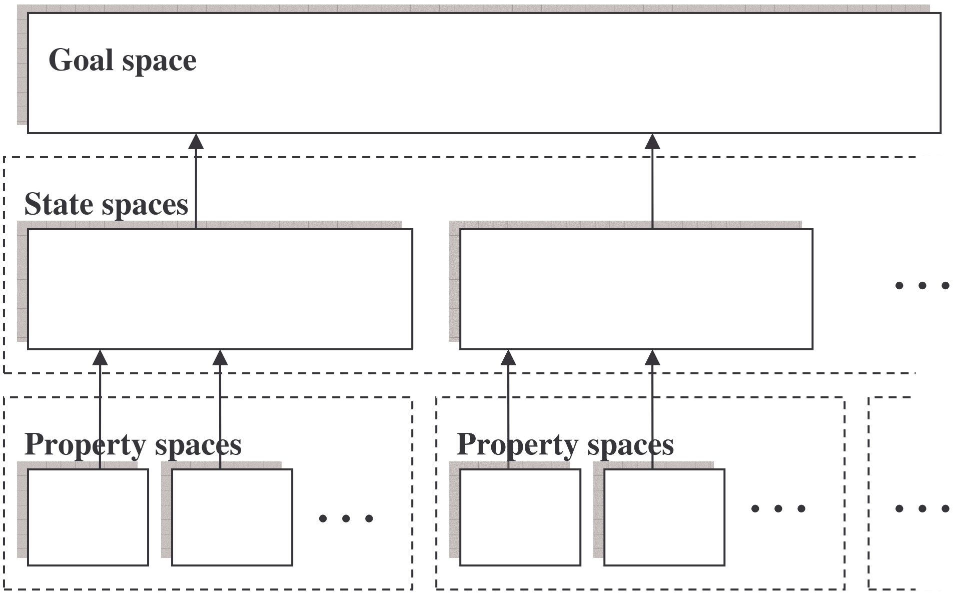


Figure 26. Goal space is a compound conceptual space with state spaces or state change spaces as domains. The state spaces in turn consist of property spaces.

It is time for an example. Consider once again the smith, who features in many examples in this thesis, who hits a piece of iron with a hammer to flatten it. I will translate the goal of the smith to a concept in goal space. Consider that the smith represents the goals as state changes in this particular case. First I define the frame terms of the property spaces involved. I will use simple integers instead of real intervals just for clarity. The first part of the goal, the change in the *position* property of the hand, can be represented as follows:

$$DROP_{\overrightarrow{pos}} = \begin{bmatrix} x:5 \\ y:0 \\ z:-10 \end{bmatrix} \tag{3.57}$$

Also his hand should turn during the drop, to make a really good swing (e.g. a change in the angle of the z axis of the hand, while the other angles remain aligned):

$$TURN_{\overrightarrow{or}} = \begin{bmatrix} \alpha:0 \\ \beta:-10 \\ \gamma:0 \end{bmatrix} \tag{3.58}$$

The entire intended state change for the smith's right hand is thus:

$$HIT_{\overrightarrow{state}} = \begin{bmatrix} \overrightarrow{pos}:DROP \\ \overrightarrow{or}:TURN \end{bmatrix} \tag{3.59}$$

Because I used state change concepts here, the concept *HIT* can be used to represent both the intended state change of the hand and the intended state change of the hammer, regardless of their different starting positions.

The iron bar is supposed to flatten, which can be represented by a point in global shape change space with a negative value for height. The intended state change of the iron is then given generically by:

$$FLATTEN_{\overrightarrow{state}} = \left[\overrightarrow{gs} : \left[a_y : \mathbb{R}_- \right] \right] \tag{3.60}$$

Now, the three concepts that describe the state changes of hand, hammer and iron can be combined to create the concept corresponding to the smith's goal. After applying these concepts to the objects involved, the following frame term is obtained:

$$FLATTENIRON_{goal}(smith) = \begin{bmatrix} \overrightarrow{state} : HIT(righthand) \\ \overrightarrow{state} : HIT(hammer) \\ \overrightarrow{state} : FLATTEN(iron) \end{bmatrix} \tag{3.61}$$

Of course, iron may be flattened in numerous other ways. The smith may be left-handed, or prefer a large rock or a press over a hammer, so the concept is actually too narrow for the concept's name. However, let us assume that the smith in this example knows no other ways to flatten iron, and always associates flattening the iron with this specific instantiation of the goal. Otherwise, naming concepts becomes a burden.

In chapter 2 I mentioned that goal representations are, just like motor representations, effector-independent. Similarly to the concepts in motor space, concepts in goal space are regarded as functions that require an object argument. Therefore, the effector-independency of goal space concepts is accounted for in a similar fashion. In chapter 2 I also claimed that goals such as in the example above should be represented as causal chains. Strictly speaking, this would require some way to order the concepts in goal space. Assuming that this is possible by adding some kind of an ordering relation to the logic arsenal, and avoiding going too deeply into the details, I will omit such an explicit ordering here. This absence will become apparent in 4.2, when the aspectual and causal structure of events are investigated.

3.4 Action space

It is time to summarize the work so far. I have formalized two high-level conceptual spaces. *Motor space*, in which entire motor patterns can be represented, consists of *segment spaces*, in which motor segments are represented. A motor segment is a motor pattern of several effector systems simultaneously, that are each represented as foxels in an *effector space*, consisting of *joint spaces* in which a constant force on a joint is represented. The other high level space is *goal space*, which supervenes on *state* and *state change spaces* that represent the endstate or state change of an object. Those spaces in turn consist of *property (change) spaces* in which properties are represented such as position, shape and temperature.

As I argued in chapter 2, an action representation contains both a motor representation and a goal representation. For that reason, I create yet another higher-level conceptual space, *action space*, which is composed of motor space and goal space. This conceptual space is illustrated in Figure 27.

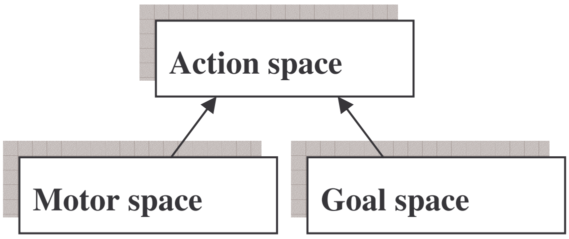


Figure 27. Action space is a compound conceptual space containing motor space and goal space.

The generic action space concept is given in (3.62) (not fully specified down to the lowest level).

$$CONCEPT_{action} = \begin{bmatrix} motor: \begin{bmatrix} segment:... \\ segment:... \\ \vdots \end{bmatrix} \\ goal: \begin{bmatrix} state:... \\ \overrightarrow{state}:... \\ \vdots \end{bmatrix} \end{bmatrix} \qquad (3.62)$$

This composition is fairly trivial, and in the next chapter some examples of action space concepts will be given. The remainder of this section is devoted to finding an appropriate similarity measure for this compound conceptual space. The hierarchical, compositional structure that gradually took shape throughout this chapter seems to have little in common with the simple, elegant conceptual spaces approach presented at the beginning. This complexity is a result of the intrinsic complexity of our actions, the great variability of motor patterns and the infinite number of intentions possible. Fortunately, the modular approach has all the merits of the conceptual spaces framework, as long as the notion of similarity is incorporated in a suitable way. Geuder and Weisgerber (2005) give an account of a similarity measure σ of two points φ and ψ that can be restated in terms of the modular conceptual space logic described here:

$$\sigma_{A\cap B}(\varphi_A,\psi_B)=w_1\cdot\sigma_{D_1}\left(D_1(\varphi_A),D_1(\psi_B)\right)\otimes\ldots\otimes w_n\cdot\sigma_{D_n}\left(D_n(\varphi_A),D_n(\psi_B)\right) \tag{3.63}$$

Here obviously all $D_i\in A\cap B$. It is left unspecified what kind of conjunction the operator $\otimes$ is. What this formula says is that the similarity of two concepts in (possibly different) domains is the weighted sum, product, or other type of operation, of the similarities of the concepts with respect to the component domains they have in common. Thus, if two concepts have no domains in common, the similarity is zero. (Note that, as with the crux and plus operations, domains that exist at different depths in the domain hierarchy cannot ‘see’ each other and are not taken into account). The weights w_i resemble the importance of each domain in the comparison. Each weight is not constant for a given domain, but is assigned a value dependent on the typical properties of the target of comparison. This makes it possible to account for asymmetrical similarities as shown in (Rosch 1975), for example that $\sigma(PENGUIN,ROBIN)$ is greater than $\sigma(ROBIN,PENGUIN)$. Such asymmetries might be (and probably will be) discovered in the actions domain, but that is something for further research.

In addition to a similarity measure, Geuder and Weisgerber argue for the importance of a *comparability* measure, to solve the problem of ‘comparing the incomparable’. In the similarity measure, only the domains that both concepts have in common are taken into account. A comparability measure on two concepts, that starts at a value of 100%, is decreased for each domain that is not shared by two concepts. The calculation of comparability in parallel to similarity enables us to regard the similarity value as a true similarity measure only when the comparability is high enough. Using these notions of similarity, weights and comparability, a distance function can be defined on the composite conceptual space designed in this chapter. Concepts can then be regarded as real regions in a metric space – although the weights must be specified. I cannot predict with certainty what kind of regions will correspond to action concepts, e.g. whether or not they are convex, or connected. But since all the building blocks (i.e. joint spaces and property spaces) can be grounded fairly directly in perception, I do not see why the line of reasoning behind convexity in color space (Warglien and Gärdenfors 2007, Jäger and Van Rooij 2007) would not apply to action space. The compound action space has all the merits of a simple conceptual space, plus modularity, flexibility, expressivity and the possible asymmetry of comparison.

4 From action space to language

This chapter shows how several phenomena in natural language might stem from our cognitive architecture, modeled by the conceptual spaces formalized in chapter 3. Three phenomena will be discussed. Section 4.1 discusses the manner/result dichotomy in verb meaning. Section 4.2 treats aspectual and causal structure in event representations as modeled by Croft (to appear). Section 4.3 discusses the semantics of modals, based on the analysis in (Talmy 2000). Section 4.4 provides a summary.

4.1 The manner-result dichotomy

Different verbs can describe different aspects of the same event. Consider a vandal hitting a window with a rock, breaking it. This event can be described in two ways, using the verbs “break” and “hit”, as in (4.1) (taken from (Levin 2007)).

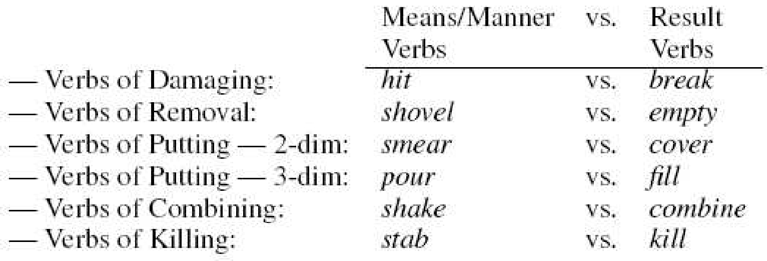
a. The vandal broke the window with a rock. (4.1)
b. The vandal hit the window with a rock.

“Break” describes a change in the state of the window, without describing the manner in which it happened. The verb “hit”, on the other hand, describes only the manner (e.g. forceful contact) and not the resulting state for the window. Not only do these verbs describe different parts of the event, they also behave differently, for example with respect to their ability for *causative alternation* as in (4.2) (taken from (Levin 2007), for a full overview of the differences in behavior, see also (Fillmore 1970)).

a. The boy broke the window.
The window broke. (4.2)
b. The boy hit the window.
* The window hit.

The differences between “break” and “hit” are not unique. Table 1 shows that the dichotomy generalizes to a wide range of examples.

Table 1. The manner/result dichotomy. Taken from Levin 2007.



	Means/Manner Verbs	vs.	Result Verbs
— Verbs of Damaging:	*hit*	vs.	*break*
— Verbs of Removal:	*shovel*	vs.	*empty*
— Verbs of Putting — 2-dim:	*smear*	vs.	*cover*
— Verbs of Putting — 3-dim:	*pour*	vs.	*fill*
— Verbs of Combining:	*shake*	vs.	*combine*
— Verbs of Killing:	*stab*	vs.	*kill*

The dichotomy is not only apparent in the behavioral patterns shown by Fillmore (1970). Gentner (1978) was one of the first to conclude from experiments that children more readily learn manner verbs than result verbs. Forbes and Farrar (1995) evaluate this field of research and come to similar conclusions, revealing a manner-over-result bias. The conceptual space defined by Geuder and Weisgerber (2002) to capture the semantics of verbs of vertical movement consists of the quality dimensions *manner* and *direction*, showing a similar distinction. Talmy (2000) noted that manner verbs are often implicitly paired with result verbs, because the two together constitute a larger causal interaction. This is visible in the table: regardless of the differences of “hit” and “break”, “shovel” and “empty”, “smear” and “cover”, each pair of verbs does seem intertwined somehow. “Hit” and “break” are both verbs of damaging, and similarly other pairs are verbs of removal, putting, combining and killing. This pairing should be taken with a grain of salt. There are obviously multiple manner verbs that can lead to a kill (“stab”, “shoot”, “hit”), and there are multiple ways to combine stuff (“shake”, “stir”), nor does each manner verb necessarily lead to the result it is paired with. But that there is some kind of intuitive pairing cannot be denied.

Manner verbs and result verbs describe different parts of the same action, and this dichotomy can be explained by looking at the representation of actions. In chapter 3 I constructed action space from motor space and goal space (Figure 27). It should be obvious that what is represented in motor space is precisely the manner of an action, and what is represented in goal space is the (intended) result. The meanings of manner verbs can thus be represented by some restricted region in motor space and the universal concept in goal space, and vice versa for result verbs, as exemplified by (4.3).

$$HIT_{action} = \begin{bmatrix} motor : DROPFAST \\ goal : \top \end{bmatrix} \qquad BREAK_{action} = \begin{bmatrix} motor : \top \\ goal : BREAKOBJECT \end{bmatrix} \tag{4.3}$$

Here the precise meaning of the concepts *DROPFAST* and *BREAKOBJECT* is left implicit. Verbs may be combined to form new concepts. For example, one could hit in a shoveling way, or one could intend to simultaneously break and kill someone. When the two verbs that are combined are a manner and a result verb from the same type of causal interaction, like hit and break, the resulting concept describes an action more completely. Using the concepts *HIT* and *BREAK*, a complete description of the action such as “the vandal smashed the window” can be obtained, as in (4.4).

$$SMASH_{action} = HIT_{action} \times BREAK_{action} = \begin{bmatrix} motor : DROPFAST \\ goal : BREAKOBJECT \end{bmatrix} \quad (4.4)$$

To summarize, the syntactic and semantic dichotomy between manner verbs and result verbs is directly reflected in the nature of action representations, consisting of a goal representation and a motor representation.

4.2 Aspectual and causal structure

The structure of events is twofold. Events are structured *causally*, bigger events being composed of smaller events that causally interact. In addition, events have an *aspectual* structure, also called lexical aspect, that determines whether an event is to be regarded as an activity, a state, an achievement, etcetera. Croft (to appear) introduces a three-dimensional representation schema for the aspectual and causal structure of events. The building blocks are two-dimensional diagrams as in Figure 28, representing the change of a certain quality of an object through time.

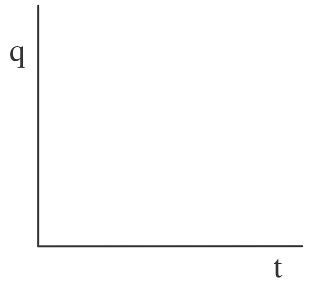


Figure 28. The building block for argument linking.

The change of a quality through time is represented in a number of phases that together form the aspectual contour $q(t)$. Different event types correspond a different shapes of the aspectual contour, but also to a different profiling of the phases in $q(t)$. I will come back to profiling later. There are three properties that contribute to the classification of events: punctual/durative, stative/dynamic and telic/atelic. Punctual event phases are points on t, durative event phases are extended on t. Stative phases are points on q, dynamic phases are extended on q. Telicity is marked by the presence of a clear result state in the diagram. I will illustrate this in more detail below.

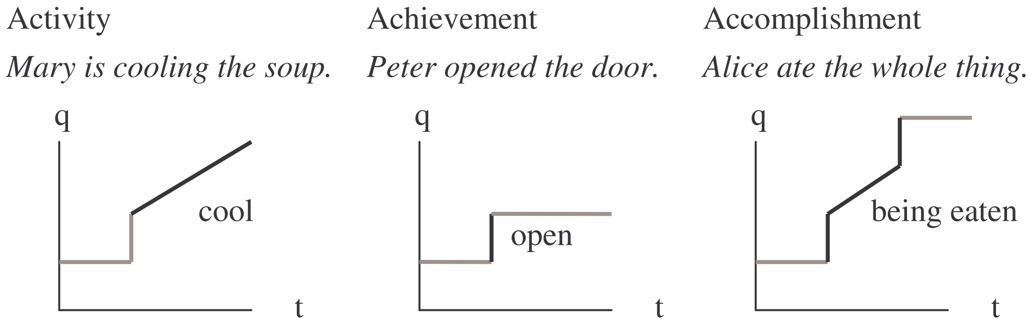
Croft distinguishes ten event types, among which are three types of *states*: inherent states (“She’s tall”), transitory states (“The window is open”) and point states (“It’s 5pm”). States are not dynamic (i.e. stative), not telic (i.e. unbounded), and durative. *Activities* are the dynamic counterparts of states. They are unbounded and durative as well. Croft discerns between directed activities ("The soup is cooling") and undirected activities ("She's dancing"). *Achievements* are dynamic as well, but telic and punctual. Croft distinguishes three kinds: reversible achievements ("The door opened"), irreversible achievements ("The window shattered") and cyclic achievements ("The light flashed"). The first two are both instances of directed achievements, which denote (punctual) transitions to transitory and inherent states respectively. Finally, Croft discerns between two kinds of *performances*, the accomplishment ("I ate the whole thing") and the runup achievement ("Help! He's dying!"). These two add the notion of telicity, boundedness, to both kinds of activities, and are the durative counterparts of achievements. The four main types are given in Table 2, where a cross indicates the presence of the property.

Table 2. An overview of the properties of four main event types.

	Dynamic	**Telic**	**Durative**
State			x
Activity	x		x
Achievement	x	x	
Accomplishment	x	x	x

Events that involve a human agent doing something correspond to actions, and therefore such event descriptions can be represented as concepts in action space. If a human agent is absent in the event description, this connection is harder to see (e.g. what kind of human action is described by "the door opened"?). It seems that events without an explicit sentient agent describe only the result of some action which is left implicit. To avoid confusion, a human agent will be made explicit in each example below (e.g. "Mary opened the door." instead of "The door opened."). The key idea is that the basic distinctions between event types are not arbitrary but are grounded in the structure of goal space. I will show this for event descriptions involving a human agent, but it applies naturally to descriptions lacking a human agent.

States have nothing to do with actions, because states are stative by definition. I will therefore only investigate the three dynamic types of events: activities, achievements and accomplishments, the aspectual contours of which are depicted in Figure 29. Activities (Figure 29 left) are atelic, i.e. in "Mary is cooling the soup" there is no clear endstate. The cooling may go on forever (practically). Therefore the goal of the event can be represented not as an endstate, but as a state change. Equation (4.5) gives the goal space concept corresponding to this.



Activity *Mary is cooling the soup.* Achievement *Peter opened the door.* Accomplishment *Alice ate the whole thing.*

Figure 29. Aspectual contours of an activity, an achievement and an accomplishment. Adapted from (Croft to appear).

$$COOL_{goal} = \left[\overrightarrow{state} : \left[\overrightarrow{tmp} : \left[t : \mathbb{R}_{-} \right] \right] \right] \tag{4.5}$$

The arrow superscripts indicate that these concepts describe a state change, as introduced in chapter 3. Since no particular manner of the event is specified (e.g. whether Mary is blowing on it or stirring it), the action described by the sentence “Mary is cooling the soup” can be represented as follows, with a universal motor space concept:

$$COOLDOWN_{action}(Mary, soup) = \begin{bmatrix} motor : \top(Mary) \\ goal : COOL(soup) \end{bmatrix} \tag{4.6}$$

Achievements (Figure 29 middle) are punctual and bounded. In “Peter opened the door”, there is a clear endstate (e.g. when the door is open), but the trajectory towards it is not specified, i.e. it is described as happening instantaneously, so here we see the opposite of an activity. Achievements can thus be represented as an endstate in goal space:

$$OPEN_{goal} = \left[state : \left[or : \left[\alpha : \pi \right] \right] \right] \tag{4.7}$$

If we now take the sentence “Peter *kicked* the door open”, which has exactly the same aspectual contour, we can create a more complete representation of the event, as in (4.8).

$$KICKOPEN_{action}(Peter, door) = \begin{bmatrix} motor : KICK(Peter) \\ goal : OPEN(door) \end{bmatrix} \tag{4.8}$$

Accomplishments, like “Alice ate the whole thing” (Figure 29 right), are the durative counterparts of achievements. In accomplishments, both an endstate is specified (having eaten the whole thing) and the trajectory towards it (eating the thing). Therefore the goal corresponding to the event of “Mary eating the whole thing” is the following twofold concept, which consists of an endstate description and a state change description (where being eaten is, for simplicity, described as a change in the wholeness dimension of material structure space):

$$EATWHOLE_{goal} = \begin{bmatrix} state : \left[ms : \left[w : 0 \right] \right] \\ \overrightarrow{state} : \left[\overrightarrow{ms} : \left[w : \mathbb{R}_{-} \right] \right] \end{bmatrix} \tag{4.9}$$

Arguably "eat" is neither a manner nor a result verb, so the sentence "Alice ate the whole thing" is a concept with a non-universal motor concept as well as a specified goal:

$$EATWHOLETHING(Alice, thing)_{action} = \begin{bmatrix} motor : FEED + CHEW(Alice) \\ goal : EATWHOLE(thing) \end{bmatrix} \tag{4.10}$$

To summarize the findings, let me expand Table 2 by adding two columns for the two types of goal representation, disregarding the stative event, as in Table 3.

Table 3. An overview of the properties of the three dynamic event types.

	Dynamic	Telic	Durative	Endstate	Trajectory
Activity	x		x		x
Achievement	x	x		x	
Accomplishment	x	x	x	x	x

As can be seen from the table, all telic events correspond to the endstate type of goal representation. All durative events correspond to the trajectory type of goal representation. Accomplishments combine these two. Now, this correspondence between event types and the structure of goal space might not be especially staggering. However, when the building blocks are, as Croft proposes, linked in a third dimension representing causality the similarities to the goal representation framework become undeniably strong. Figure 30 represents the causal and aspectual structure of the sentence "Sally removed the book from the table".

This structure is three-dimensional, but for display purposes the third dimension, causality, is represented by the vertical arrows. In the figure, Sally's manipulation *causes* the book to move. During the event the table is present but not engaged in any causal relationship (the book and the table in the figure are not connected by arrows), so I will omit the table here. This event consists of an undirected activity (manipulate) and an accomplishment (move from the table). This corresponds to the conceptual goal-space representation in (4.11).

$$REMOVEFROMTABLE_{goal}(Sally, book) = \begin{bmatrix} \overrightarrow{state} : NOTIDLE(Sally) \\ \overrightarrow{state} : MOVE(book) \\ state : FROMTABLE(book) \end{bmatrix} \tag{4.11}$$

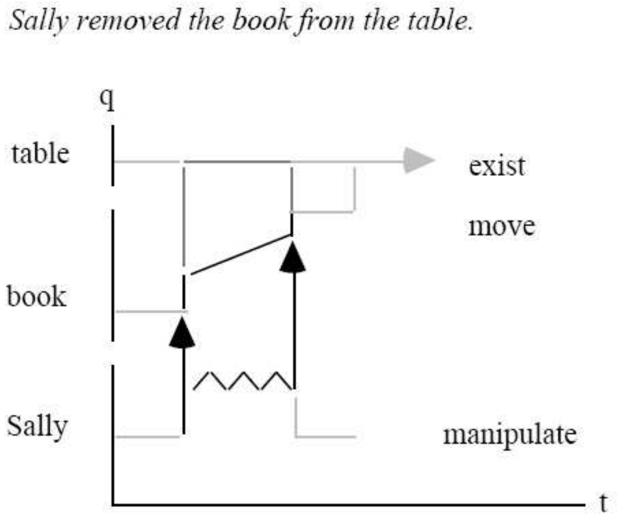


Figure 30. A three-dimensional representation of causal and aspectual structure. From (Croft to appear).

Currently, there are two entries in the concept for the book object, one for the state change and one for the endstate. Goal space thus *conflates* causal and aspectual structure, which is exactly what Croft's approach is meant to solve. However, as I mentioned in chapter 3, some kind of an ordering relation could be added to the concepts in goal space to account for it. I will not go into this here.

Equation (4.11) gives the representation of the entire scene. According to Langacker (1987), a verb in a particular context denotes or *profiles* one or several phases of the event. In Croft's diagrams, the parts of the contour that are profiled by the verb are drawn in black. This profiling is apparent in the simple diagrams of Figure 29, but also in the more complex sentence "Sally removed the book from the table" of Figure 30. In the latter sentence, the profiled part of the verb "remove" is restricted to Sally manipulating and the book changing state (including endstate). At a conceptual level, the distinction can also be made. The profiled part of the verb "remove" in this context is given in (4.12).

$$REMOVE_{goal}(Sally, book) = \begin{bmatrix} \overrightarrow{state} : NOTIDLE(Sally) \\ \overrightarrow{state} : MOVE(book) \end{bmatrix} \tag{4.12}$$

I have only discussed the three main dynamic event types. The subtle differences between e.g. directed and undirected activities, or reversible and irreversible achievements remain to be explained. Such differentiation might follow from background knowledge about the domain, for example about the reversibility of certain state changes. Breaking is intrinsically irreversible, so breaking the glass is an irreversible achievement. Undirected activities may be characterized by some concept $NOTIDLE$, as in (4.11), indicating that she does something but it does not matter what. This concept is the universal concept $\top$ minus everything at or near zero change. The cyclicity of some achievements cannot be accounted for directly in the current framework, but they could instead be regarded as the repetition of a single achievement.

4.3 The greater modal system

Talmy devotes a chapter of his book "Toward a Cognitive Semantics" (2000) to force dynamics in language and cognition. He analyses the meaning of words such as "cause", "let", "help" and "despite" and presents a way of diagramming those meanings. According to Talmy, the meaning of such verbs is grounded in an interaction between an *agonist*, the focal force entity, and an *antagonist*, the force opposing it. It is the agonist's tendency, the balance of forces and the resultant force that determine the precise meaning. I will briefly introduce the diagramming by means of the four basic examples in Figure 31.

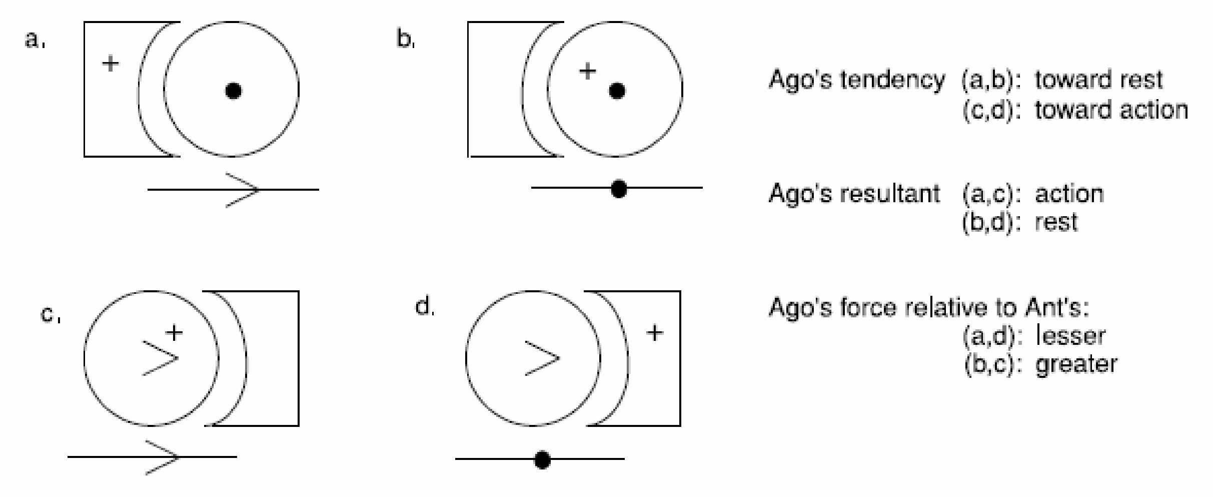

a.
+
b.
+
c.
+
d.
+
Ago's tendency (a,b): toward rest
(c,d): toward action
Ago's resultant (a,c): action
(b,d): rest
Ago's force relative to Ant's:
(a,d): lesser
(b,c): greater

Figure 31. The basic steady-state force-dynamic patterns. Taken from (Talmy 2000).

In the diagrams, the circle denotes the agonist and the socket denotes the antagonist. The plus sign denotes which party is the strongest. The bullets and angle-brackets in the agonist denote the agonist's tendency towards rest and towards action respectively. The same signs on the line segment beneath each diagram denote the agonist's resultant, which is either rest or action. The four diagrams in the figure are the underlying force-dynamic patterns of the sentences in (4.13), taken from (Talmy 2000).

a. The ball kept rolling because of the wind blowing on it.
b. The shed kept standing despite the gale wind blowing against it. (4.13)
c. The ball kept rolling despite the stiff grass.
d. The log kept lying on the incline because of the ridge there.

Now, the resultant is in fact redundant: it can be determined from the balance of forces, assuming, like Talmy, that the antagonist always opposes the agonist's tendency. That leaves us with the following basic ingredients for Talmy's force-dynamic approach: two entities (let's stick with the terms "agonist" and "antagonist"), the agonist's tendency, and the balance of forces.

Consider two human agents. The actions of these two agents can naturally be represented as concepts in the action space formalized in chapter 3. A person, say John, who is running with an intention or tendency to keep running, corresponds to roughly the following concept:

$$RUN_{action}(John) = \begin{bmatrix} motor: RUNNING(John) \\ goal: \left[\overrightarrow{state}: FORWARD(John)\right] \end{bmatrix} \tag{4.14}$$

Another person, say Mary, might be trying to block John, which can be represented as:

$$BLOCK_{action}(Mary, John) = \begin{bmatrix} motor: BLOCKING(Mary) \\ goal: \left[state: STOP(John)\right] \end{bmatrix} \tag{4.15}$$

The interaction of both persons' actions can be represented in a conceptual space that is supervening on two action spaces, which I will call *interaction space*, as illustrated in Figure 32.

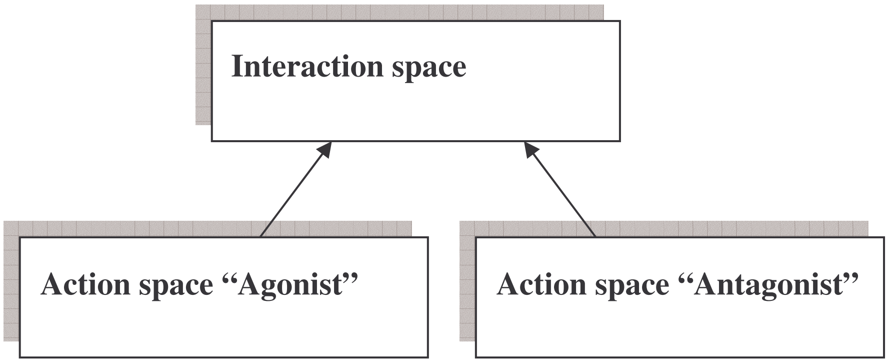


Figure 32. Interaction space as a compound space of two action spaces.

For convenience I have given both action spaces different names. The force-dynamic interaction can now be represented as the concept in interaction space given in (4.16).

$$PREVENT_{interaction}(Mary, John) = RUN_{ago}(John) \wedge BLOCK_{ant}(Mary) \tag{4.16}$$

Of course, an assumption needs to be made concerning which effector of John is interacting with which effector of Mary, but in principle this representation contains each of Talmy's ingredients. There are two entities (John and Mary), there is the tendency of John

(*FORWARD*), and the balance of forces is available (we can simply compare the resultant force of *RUN* with that of *BLOCK*). From this, also the resulting action can be derived.

It is important to stop for a moment and consider the difference between the representation of an interaction in this double action space and the representation of a sentence like “Mary removed the book from the table” in goal space, as was done in the previous section. In the latter case, only the intention of Mary is recorded, without any true force dynamics. To make it more complete, the motor patterns of Mary removing the book can be added as a motor space concept, which then combines with the goal space concept to represent the entire action. However, this full representation still lacks a representation of the forces exerted by and the tendency of the antagonist (i.e. the book). The meanings of regular verbs are fully captured by an intention (goal space) and motor patterns (motor space), whereas the modals require *two* of these. For example, to understand “Mary removed the book from the table”, you do not need to know the book’s tendency or the forces exerted by it. It is enough to picture Mary and her intention. However, to understand the meaning of “Mary *should* (or *can, must, ...*) remove the book from the table”, knowing only Mary’s intention and motor patterns is not enough.

The action space formalized in this thesis is meant only for human physical actions, but it should be apparent that, given an ‘action space’ appropriate for rolling balls and lying logs, the approach sketched here naturally extends to Talmy’s examples. Talmy gradually increases the expressivity of his framework by adding dynamicity, psychosocial powers and foregrounding differences to the four basic types in Figure 31. The full-blown framework can account for the meaning of all verbs in what Talmy calls the *greater modal system*. These are verbs like “can”, “may”, “must”, “should”, “dare”, and “ought”. These verbs share many strange grammatical features, such as the lack of “to” for the infinitive form of the following verb, the lack of “-s” for the third-person singular, postposed “not”, and inversion with the subject in questions (Talmy 2000). I claim that what makes these verbs special is that they are of a higher degree compared to regular verbs, because their meanings supervene on a double action space. I have already shown this for the four basic force-dynamic patterns in Figure 31. What remains is to show that also the more complex features in Talmy’s framework are already or can be incorporated in my framework.

The dynamicity Talmy adds to his framework through features like a shift in state of impingement and a shift in balance of strength is already present in my framework, in the sense of *motor segments*. In motor space, any complex, dynamically changing motor pattern can be represented. Talmy furthermore extends his framework to include psychological and social forces and tendencies. I have already briefly mentioned in chapter 3 that the modular conceptual spaces approach can easily be extended to include such notions. This is of course very useful to represent events like "he held himself back from responding" or "the man resisted the pressure of the crowd against him", which are purely psychological or social. However, Talmy also makes a great deal of the mental states of sentient agents when they engage in otherwise purely physical force interactions, e.g. in "the attendant restrained the patient", in which case I think there is no added value in a whole representation of the conflicts in the inner selves of the agents. In any case, the structure of goal space, when expanded with psychosocial property spaces, allows the representation of purely psychological or social intentions of both the agonist and antagonist even for physical events.

So, the framework, with some extensions, can account for the meanings of modals as proposed in (Talmy 2000). In fact, applying the framework to the modal verbs gives some insight in the relation between the modals and regular verbs. Modals are of a higher conceptual level. Why this conceptual difference leads in particular to the syntactical differences described by Talmy is a question for further research.

4.4 Summary

In this chapter I have shown how three major linguistic phenomena can be explained in terms of the action space formalized in chapter 3. First, the manner-result dichotomy in verb meaning is reflected by the bipartition of action space into motor space and goal space. Second, I have shown how different aspectual event types correspond to different goal space representations of the action involved. Third, I have shown that what makes modal verbs special is that their meanings are of a higher level compared to regular verbs. Modal verbs denote regions in interaction space, a conceptual space supervening on two action spaces.

5 Conclusions and outlook

In this chapter I will look back on the aims of this thesis, give a brief summary and draw several conclusions. A few interesting suggestions for further research are given, both for the area of robotics and for cognitive science.

5.1 Conclusions

The aims of this thesis were to construct a model of human action representations, to construct it in such a way that it would in principle be implementable, and to explain three linguistic phenomena in terms of this model. To these ends, existing approaches towards action representation have been compared and the conceptual spaces approach was chosen as a starting point. A modular conceptual space logic was formalized and a hierarchical, composite conceptual space for the representation of actions was constructed. Three linguistic phenomena have been tackled. The manner-result dichotomy in verb meaning was shown to be reflected by the motor-goal partition of action space. Aspectual event structure was grounded in the structure of goal representations according to the independent control model. The special status of the modal verbs was confirmed by the fact that their meanings supervene on two regular action spaces.

The modular conceptual space that was constructed is psychologically realistic in the following ways. The conceptualization is hierarchical and the lowest levels for both goal representation (properties) and motor representation (joints) are very simple and plausible building blocks. The independent control model is applicable to goal space. Effector systems play a central role in action space, and the concepts for motor representations as well as goal representations behave in an effector-independent way. The model incorporates the joint redundancy present in the human body. Furthermore, action space is in its current form expressive enough for virtually all physical actions involving a human agent and one or more tools and objects. Besides that, due to the modularity it can fairly easily be extended to include other types of actions. Compared to existing (conceptual, symbolic and associationist) approaches, I think it is safe to assert that the model presented in this thesis is superior in both expressivity and cognitive realism.

One question can be asked concerning the discriminativity of the model, especially the motor part. In designing the joint spaces, I tried to stick as closely as possible to the signals that would have to be sent to the muscles in order for some movement to take place. I did not take into account the joint angles themselves. This raises the question whether the model can discern between, for example, *catching* some heavy object and *throwing* some heavy object. Both movements seem similar – the arms apply a force to the object from underneath – but in throwing the object, force and motion are collinear (i.e. the object accelerates), whereas in catching the object force and motion are not collinear (i.e. the object decelerates). It seems that the model misses some important information by including only the forces. To account for the motoric differences between e.g. catching and throwing, an extra dimension ‘initial angle’ could be added to each joint space. The human proprioceptive system is in principle able to gain information about joint angles, but whether this information is used in the motor cortex remains to be investigated.

The three linguistic phenomena have been accounted for quite successfully. The only shortcoming occurred in capturing the causal structure of events. The model conflates causal and aspectual structure in goal space. This is due to the fact that goal space does not come with a mechanism to represent the direction of causality. States and state changes of different objects are all represented as equals, whereas to represent the causal structure, the states and state changes have to be ordered somehow. In constructing the model, I have avoided this burden just as I have avoided making explicit the order of motor segments in motor space. Still, I think that it is only a minor addition to add this kind of structure, e.g. an ordering relation, to a conceptual space.

Whether the model, aside from being psychologically realistic and linguistically explanatory, is also computationally powerful enough is not trivial. Several applications in robotics and motion recognition make use of conceptual spaces (e.g. Chella, *et al.* 2000), so it is certainly possible, but it seems that those conceptual spaces are far simpler than the massive, hierarchical structures presented in this thesis. To be honest, the model presented here seems rather extensive and complex for only relatively simple actions. Fortunately, the model is very well structured compared to the conceptual spaces used in existing approaches and the building blocks (joint spaces, property spaces) are only of low dimensionality. The rich structure can be exploited computationally by an attention mechanism that focuses on just a selected subset of the building blocks, expanding the focus only when the action cannot be reliably classified. In general I have good hopes that the model will provide a very flexible yet powerful tool for applications in robotics.

Overall the goals of this thesis have been attained. In general, insight has been gained in action representations and the bridge between cognition and language has been strengthened. Last but not least, the conceptual spaces approach has gained some more credibility.

5.2 Outlook

The ultimate test of the model would be to investigate human similarity judgments in the actions domain and compare the results with the predictions made by the model. This would require a lot of participants to judge a large set of action pairs, possibly on film. Ideally, the similarity judgments would correspond to the similarity measure of Geuder and Weisgerber (2005) adopted here. The particular setting of the weights in the similarity measure that corresponds the most to human judgments should be reflected in some interesting aspects of human cognition. As I mentioned in chapter 3, an advantage of the weighted modular approach is that it can capture the asymmetry of similarity judgments. Analogous to the robin/penguin example given, I predict that hopping will be judged to be more similar to walking (the most important property of the target of comparison being forward motion) than vice versa (the most important property being one-leggedness).

The similarity measure provides a foundation for several other hypotheses for further research. In chapter 2 I mentioned that, as a task becomes more accustomed, the patient (and instrument) parts of goal representations somehow seem to become less important and eventually disappear. It seems that the weights assigned to the instrument and patient domains become smaller and smaller as one gets more accustomed to a task. This explanation would have to be apparent in the similarity judgments of people with different skill levels. It is an empirical issue whether this is indeed the case.

On a higher level, the weights might be responsible for linguistic fine-tuning of the evolved cognitive machinery. According to the *Sapir-Whorf hypothesis*, differences in language cause differences in the perception and cognition of speakers:

> ... the ‘real world’ is to a large extent built up on the language habits of the group... We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation. (Sapir 1929).

Most language differences are too young to have originated from structural differences in the evolution of the deeper cognitive architecture of speakers. But language differences may lead to a different weights setting in a speaker’s conceptual spaces, which would affect the speaker's similarity judgments. The following is meant to exemplify this line of thought, and it is highly hypothetical. The way in which the *path* of an action is encoded is language-dependent, as was noted by Talmy (1985). In *satellite-framed* languages like English, the verb is generally used to express the manner and a satellite is used to express the path (e.g. "He ran into the house"). This is the case for most of the English verbs, with the exception of Latinate verbs like "exit" and "ascend". *Verb-framed* languages like Spanish, a more direct descendant of Latin, work the other way around, describing actions usually like "Entró corriendo" ("He entered running"). Although there is no difference between these languages as far as the richness of meaning is concerned, the meaning is located in different parts of the sentence, resulting in a difference in the path's and manner's *importance* for sentence meaning. A similar case was made by Slobin (1996), who compared Spanish and English with respect to manner, direction, location and trajectory. Levin and Rappaport Hovav (1992) take path to be a type of result, so the manner/path distinction is related to the manner/result distinction that is so apparent in action space. It is likely therefore that the type of verb-framing in a language would for a native speaker affect the weights assigned to goal space and motor space. Speakers of verb-framed languages such as Spanish, in which path is encoded in the verb, might base their categorization of actions more on path than on manner. This prediction, too, can be verified by investigating similarity judgments.

A related issue has to do with *foregrounding* in Talmy's (2000) approach towards the meaning of modals. Talmy presents two foregrounding mechanisms. The first is achieved through foregrounding either the agonist (e.g. "the ball is rolling because of the wind") or the antagonist ("the wind is making the ball roll"). The second foregrounding mechanism involves naming either the resultant ("the shutting valve made the gas stay inside") or the tendency ("the shutting valve stopped the gas from flowing out"). Foregrounding does not seem to really affect the meaning of a sentence – both sentences are still a description of the same event. Therefore the foregrounding differences cannot be accounted for by associating a different concept with each sentence. Instead, foregrounding might be a result of changing the weights of interaction space. By foregrounding the agonist, it becomes more important in a comparison. Compare the sentences in (5.1), and those in (5.2). Both describe the same pair of events.

a. The ball is rolling because of the wind.
b. The ball is rolling because of the slope. (5.1)

a. The wind is making the ball roll.
b. The slope is making the ball roll. (5.2)

The scenes all have the same agonist (the ball) but different antagonists (the wind and the slope). I predict that the events described by the sentences in (5.1), with a foregrounded agonist, are judged to be more similar than the events described in (5.2), with a foregrounded antagonist. Similar pairs of sentences could be given for resultant/tendency foregrounding, leading to a similar prediction. Whether foregrounding is really a matter of weight-adjustment remains a question for further research, but it seems plausible.

Doing research on similarity judgments may reveal that the current conceptual framework is lacking some dimensions. For example, it might lack a certain degree of proprioception as mentioned in the previous section. This research could also reveal that in fact the resultant force on the effector is all that matters, rendering the joint spaces worthless. In any case, after an adequate similarity measure has been established, it will be particularly interesting to look at the concepts themselves. To what kind of regions do action concepts correspond? Are they convex, or maybe only connected? Is a notion similar to that of 'natural concepts' applicable to the actions domain, and do such 'natural actions' correspond to convex regions in action space? It seems to me that surely the simpler kinds of modals, e.g. the ones which describe the four situations depicted in Figure 31, should display some kind of convexity in their interaction space. Gärdenfors (2007) has suggested that functional properties or *affordances* (e.g. whether you can throw it, eat it, walk on it, hit something with it), which seem essential for the categorization of tool-like objects, should correspond to convex regions in his theoretical action space. Whether it actually works out like this in my framework remains to be established.

On the implementational side, research needs to be done on automated force decomposition from dynamic scenes and on the recognition of the intentions of other agents, as described in chapter 2. Furthermore, to make the framework described in this thesis suitable for implementation, the computational complexity has to be reduced. An attention mechanism could be provided, possibly similar to the one in Chella, *et al* (2000). When this is done, equipping a robot with the framework could turn out to be an effective tool for example in imitation learning for humanoid robots, classification in security camera software and sign-based human-computer interaction.

It is obvious that a lot of work still needs to be done to model all the aspects of human cognition, to understand why natural language is the way it is, and to build a robot that has a human-like representation of the world. The history of science is marked by overly optimistic predictions, especially in artificial intelligence, so a little modesty is appropriate. Nevertheless, this thesis has taken a step towards the goal in each of these disciplines, and I believe these steps are promising. I will conclude with the final sentence of the first scientific article I had to read for my study: Computing Machinery and Intelligence, written by Alan Turing.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

- A.M. Turing (1950)

6 References

Abrams, R.A., Van Dillen, L., Stemmons, V.

1994 Multiple sources of spatial information for aimed limb movements. In C. Umiltfi, M. Moscovitch (Eds.), *Attention and performance XV*: 267-290. MIT Press.

Barbey, A.K., Wolff, P.

2007 Learning Causal Structure from Reasoning, In *Workshop: Causality, Mechanisms, and Psychology*, Pittsburgh, PA.

Biederman, I.

1987 Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review* **94**: 115-147.

Chella, A., Dindo, H., Infantino, I.

2006 A Cognitive Framework for Learning by Imitation, *Robotics and Autonomous Systems* **54**:403–408.

Chella, A., Frixione, M., Gaglio, S.

2000 Understanding Dynamic Scenes, *Artificial Intelligence*, **123:**89-132.

2001 Symbolic and Conceptual Representation of Dynamic Scenes: Interpreting Situation Calculus on Conceptual Spaces. F. Esposito (Ed.): *AI*IA, LNAI* **2175**:333-343.

Cisek, P., Crammond, D.J., and Kalaska, J.F.

2003 Neural activity in primary motor and dorsal premotor cortex in reaching tasks with the contralateral versus ipsilateral arm. *Journal of Neurophysiology* **89**(2): 922-942.

Croft, W.

To appear Aspectual and causal structure in event representations. In Gathercole, V. (Ed.), *Routes to language development: in honor of Melissa Bowerman*. Lawrence Erlbaum Associates.

Dutton, J.M., Starbuck, W.H.

1971 Computer Simulation of Human Behavior, Academic Press, New York.

Feldman, J., Narayanan, S.

2004 Embodied Meaning in a Neural Theory of Language. *Brain and Language* **89**:385-392.

Fillmore, C.J.

1970 The Grammar of Hitting and Breaking. In R.A.
Jacobs and P. S. Rosenbaum (Eds.), *Readings in English*
Transformational Grammar:120-133. Ginn and Company.

Fischer Nilsson, J.

1999 A Conceptual Space Logic, E. Kawaguchi, *et al.* (Eds.), *Information Modeling and Knowledge Bases XI*:26-40, IOS Press/Ohmsha.

Forbes, J.N., Farrar, M.J.

1995 Learning To Represent Word Meaning: What Initial Training Events Reveal About Children's Developing Action Verb Concepts, *Cognitive Development* **10**:1-20

Gazzaniga, M.S., Ivry, R.B., Mangun, G.R.

2002 Cognitive Neuroscience: the Biology of the Mind, 2nd edition.

Gärdenfors, P.

2007 Representing actions and functional properties in conceptual spaces, pp. 167-195 in *Body, Language and Mind, Volume 1: Embodiment*, in T. Ziemke, J. Zlatev and R. M. Frank, Mouton de Gruyter (Eds.), Berlin.

2000 Conceptual Spaces: The Geometry of Thought, MIT Press, Cambridge, MA.

Gentner, D.

1978 On relational meaning: The acquisition of verb meaning, *Child Development* **49**: 988-998.

Geuder, W. and Weisgerber, M.

2002 Verbs in Conceptual Space. In Katz, Reinhard and Reuter (Eds.), *Proceedings of SuB6*. University of Osnabrück, 81-83.

2005 On the Geometrical Representation of Concepts. Ms. Universität Konstanz.

Giese, M.A., Poggio, T.

2002 Biologically Plausible Neural Model for the Recognition of Biological Motion and Actions, *CBCL Paper* #**219**/*AI Memo* #**2002-012**, MIT Press, Cambridge, MA.

Hård, A., Sivik, L.

1981 NCS - Natural Color System: A Swedish standard for color notation. *Color Research and Application* **6**(3):129-138.

Harnad, S.

1990 The Symbol Grounding Problem, *Physica* **D 42**: 335-346.

2005 To Cognize is to Categorize: Cognition is categorization, in Lefebvre, C. and Cohen, H. (Eds.) *Handbook of Categorization,* Elsevier.

Helbig, H., Graf, M., Kiefer, M.

2004 The role of action affordances in visual object recognition, *Perception* **33** ECVP

Hernandez Cruz, J.L.

1998 Mindreading: Mental State Ascription and Cognitive Architecture, *Mind & Language* **13**(3):323-340

Hoshi, E., Tanji, J.

2000 Differential roles of neuronal activity in the supplementary and presupplementary motor areas: From information retrieval to motor planning and execution. *Journal of Neurophysiology* **92**:3482-99.

Ilg, W., Bakir, G.H., Franz, M.O., Giese, M.

2003 Hierarchical Spatio-Temporal Morphable Models for Representation of complex movements for Imitation Learning. *11th International Conference on Advanced Robotics* (2), 453-458.

Jäger, G., Van Rooij, R.

2007 Language structure: psychological and social constraints. *Synthese,* **159**(1):99-130.

Kilner, J.M., Vargas, C., Duval, S., Blakemore, S.-J., Sirigu, A.

2004 Motor activation prior to observation of a predicted movement. *Nature Neuroscience* **7**(12):1299-301

Köhler, W.

1925 The Mentality of Apes, *Harcourt Brace and World*, New York.

Kohler, E., Keysers, C., Umilt, M.A., Fogassi, L., Gallese, V., Rizzolatti, G.

2002 Hearing sounds, understanding actions: action representation in mirror neurons. *Science* **297**:846-848.

Kreutz-Delgado, K., Long, M., Seraji, H.

1992 Kinematic Analysis of 7-DOF Manipulators, *The International Journal of Robotics Research* **11.5**:469-481

Langacker, R.W.

1987 Nouns and Verbs, *Language* **63.1**:53-94

Levin, B.

2007 The Lexical Semantics of Verbs, I: Introduction and Causal Approaches to Lexical Semantic Representation, course handout, Stanford University.

Levin, B., Rappaport Hovav, M.

1992 The Lexical Semantics of Verbs of Motion: The Perspective from Unaccusativity, in I.M. Roca (Ed.), *Thematic Structure: Its Role in Grammar, Foris*:247-269, Berlin.

McCarthy, J., Hayes, P.

1969 Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie (Eds.) *Machine Intelligence*, **4**:463–502. Edinburgh University Press, 1969.

MacKay, D.G.

1987 The organization of perception and action: a theory for language and other cognitive skills, Springer-Verlag, New York.

Marr, D., Vaina, L.M.

1982 Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London* ***B*** **214**:501-24

Marr, D., Nishihara, H.K.

1978 Representation and Recognition of the Spatial Organization of Three Dimensional Shapes, *Proceedings of the Royal Society of London* ***B*** **200**:269-294.

Medler, D.A., Dawson, M.R.W.

1998 Connectionism and Cognitive Theories, *Psycoloquy*: **9**(11).

Povinelli, D.J.

2000 Folk Physics for Apes: The Chimpanzee's Theory of How The World Works. Oxford University Press.

Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L.

1996 Premotor cortex and recognition of motor actions. *Cognitive Brain Research* **3**:131–141.

Rosch, E.H.

1975 Cognitive Reference Points, *Cognitive Psychology* **7**:532-547.

Runesson, S.

1994 Perception of biological motion: The KSD-principle and the implications of a distal versus proximal approach, in Jansson, Bergström and Epstein (Eds.), *Perceiving Events and Objects*:338-405, Lawrence Erlbaum Associates.

Runesson S., Frykholm G.

1981 Visual perception of lifted weights. *Journal of Experimental Psychology: Human Perception and Performance* **7**:733-740.

Polk, T., Simen, P., Lewis, R., Freedman, E.

2002 A Computational Approach to Control in Complex Cognition. *Cognitive Brain Research* **1**:71-83.

Sapir, E.

1929 The Status Of Linguistics As A Science, *Language* **V**:209-210.

Shen, L., Alexander, G.E.

1997 Neural Correlates of a Spatial Sensory-To-Motor Transformation in Primary Motor Cortex. *Journal of Neurophysiology*: **77.3**:1171-1194

Slobin, D.I.

1996 From "thought and language" to "thinking to speaking". In Gumperz & Levinson (Eds.), *Rethinking linguistic relativity*, pp.70-96, Cambridge University Press.

Sternad, D., Schaal, D.

1999 Segmentation of endpoint trajectories does not imply segmented control. *Experimental Brain Research* **124**:118-136.

Suddendorf, T. Corballis, M.C.

1997 Mental time travel and the evolution of the human mind. *Genetic Social and General Psychology Monographs* **123**:133-167.

Talmy, L.

1985 Lexicalization patterns: Semantic structure in lexical forms. In Shopen (Ed.): 57–149, Cambridge University Press.

2000 Toward a Cognitive Semantics, Vol. 1, MIT Press, Cambridge, MA.

Tomasello, M., Call, J.

1997 Primate cognition. Oxford University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.

2005 Understanding and sharing intentions: The origins of cultural cognition, *Behavioral and brain sciences* **28**:675-735, Cambridge University Press.

Warglien, M., Gärdenfors, P.

2007 Semantics, conceptual spaces and the meeting of minds. *Journal of Philosophy.*

Wolff, P.

2007 Representing Causation, *Journal of Experimental Psychology: General.*

Wikipedia (http://en.wikipedia.org/), various authors, on the following subjects:

Multidimensional scaling

Natural Color System, Opponent process

Sapir-Whorf hypothesis

Situation Calculus

7 Appendix: Modular Conceptual Space Logic

7.1 Interpretation

$$[\![\bot]\!] = \{\} \tag{7.1}$$

$$[\![\top_A]\!] = U_A \tag{7.2}$$

$$[\![\varphi_A \wedge \ldots \wedge \psi_B]\!] = \{\langle X, \ldots, Y\rangle \mid X \in [\![\varphi_A]\!], \ldots, Y \in [\![\psi_B]\!]\} \tag{7.3}$$

$$[\![\varphi_A + \psi_A]\!] = [\![\varphi_A]\!] \cup [\![\psi_A]\!] \text{ for } |A| = 1 \tag{7.4}$$

$$[\![\varphi_A \times \psi_A]\!] = [\![\varphi_A]\!] \cap [\![\psi_A]\!] \text{ for } |A| = 1 \tag{7.5}$$

$$[\![\varphi_A \leq \psi_A]\!] = [\![\varphi_A]\!] \subseteq [\![\psi_A]\!] \text{ for } |A| = 1 \tag{7.6}$$

$$[\![a_A : \varphi_B]\!] = \{x \mid \exists y((x, y) \in a_A \wedge y \in [\![Dom(A, \varphi_B)]\!]\} \tag{7.7}$$

$$[\![Dom(B, \varphi_A)]\!] = [\![\psi_B]\!] \text{ such that } \psi_B \text{ is } \varphi_A \text{ restricted to domain } B \tag{7.8}$$

7.2 Identities

$$\varphi_A + \bot = \varphi_A \tag{7.9}$$

$$\psi_A \times \top_A = \psi_A \tag{7.10}$$

$$\varphi_A + \psi_B = \bigwedge_{D \in A \cup B} Dom(D, \varphi_A) + Dom(D, \psi_B) \tag{7.11}$$

$$\varphi_A \times \psi_B = \bigwedge_{D \in A \cup B} Dom(D, \varphi_A) \times Dom(D, \psi_B) \tag{7.12}$$

$$\varphi_A = \bigwedge_{D \in A} Dom(D, \varphi_A) \tag{7.13}$$

7.3 Notation conventions

$$B(\varphi_A) := Dom(B, \varphi_A) \tag{7.14}$$

$$\begin{bmatrix} a_1 : \varphi_1 \\ \vdots \\ a_n : \varphi_n \end{bmatrix} := a_1(\varphi_1) \times a_2(\varphi_2) \times \ldots \times a_n(\varphi_n) \text{ (frame term)} \tag{7.15}$$

$$\begin{bmatrix} D_1 : \varphi_A \\ \vdots \\ D_n : \varphi_A \end{bmatrix} := \varphi_{D_1} \wedge \ldots \wedge \varphi_{D_n} \text{ (domain term)} \tag{7.16}$$