The role of affordances in visually situated language comprehension

Craig Chambers
Department of Psychology
University of Toronto
craig.chambers@utoronto.ca

1. What is an affordance?

The term *affordance* was coined by J. J. Gibson (e.g., Gibson, 1977)—a major proponent of ecological approaches to perception. Although the specific meaning of this concept has evolved since its original inception (see Jones, 2003), a widely-accepted definition is that an affordance is a potential way of bodily interacting with an object that is evident to an animal on the basis of perception. In classic work on this topic, the most frequently discussed features of objects used to establish affordances are their surfaces, which yield information regarding size/shape properties, material, and objects' orientation relative to a perceiver. This information can be used to determine-- at least in part-- whether and how an object could be pushed or picked up with one or two hands, whether it could support the perceiver's body, and so on. However, more abstract information, including the conceptual category to which an object belongs, may also play a role-- a point that will be discussed shortly.

Although the focus of this chapter will be limited to affordances gleaned via vision (reflecting the standard approach within ecological psychology), it is worth noting that affordances can also be detected via other sensory modalities. For example, haptic information acquired through manual exploration provides an obvious way to obtain various kinds of knowledge relevant to bodily interaction, such as when one navigates a room in the dark, or when a home renovator feels out whether a rickety old ladder is likely to support his or her weight. Similarly, the specific "clunk" sound heard when an object is placed on a rigid surface can sometimes reflect the object’s mass, in turn yielding information about how an agent might interact with it.

Exactly how do animals-- and humans in particular-- apprehend the affordances of an object? This question is tied to a number of current and past debates in ecological psychology, and I will touch on only a few of the relevant issues here. One key idea is that affordances are relativized to both the agent-perceiver and the immediate environment. For instance, an object that could support the weight of a child would not necessarily hold the same affordances for an adult, nor would the same notion of support be relevant in zero gravity. Affordances therefore reflect a kind of situation-specific and complementary relation between a given organism and the environment (see Shaw & Turvey, 1981, for additional discussion). As such, affordances are in principle quite different from what we consider to be the stable (perceiver- and context-invariant) attributes of individual objects, or the prototypical semantic properties associated with abstract object concepts. Although this aspect of affordances is uncontroversial, the details regarding how this complementary relation is actually computed by an organism are not fully understood—a point that becomes especially apparent when considering that actions performed by humans are often mediated and involve tools of various sorts (see, e.g., van Leeuwen, Smitsman, & van Leeuwen, 1994).

A second issue concerns the extent to which affordances could be straightforwardly derived from sensory-perceptual information (the "ambient optic array") without some contribution from internal conceptual representations. I will illustrate this issue using an example appearing in some of Gibson's writings, namely the case of a mailbox, which is said to afford the action of mailing a letter (Gibson, 1979). The relevant observation here is that this particular affordance seems difficult to arrive at simply on the basis of the surface features of a mailbox. To be sure, some outwardly-visible properties of a mailbox are important for establishing its potential for mailing letters, such as its apparent rigidity and the capacity to contain objects of a

certain size and shape. However, these features are common to other kinds of containers and consequently do not distinguish mailboxes from things that do not afford letter mailing. To identify the relevant affordance upon perceiving the mailbox, it seems necessary for a perceiver to also draw on stored background knowledge that allows him or her to (i) identify the specific object as an exemplar of the concept "mailbox", and (ii) extend a (primary) function associated with "mailbox" to this specific exemplar.

To some degree, Gibson's use of the mailbox example in his writings is surprising because his work reflected a generally negative view toward the need for internal representations (indeed, Gibson use of this example was not to illustrate the role of higher-level ontological influences). However, some authors have suggested that Gibson's approach to affordances was broader than his approach to other perceptual phenomena (see, e.g., Greeno, 1994). Regardless, it is on this point that inconsistencies in the use of the term *affordance* are apparent in both scientific and philosophical literatures. In some cases, a strict sensory-based interpretation of the term is adopted, whereas in other cases affordances are clearly intended to reflect a perceiver's past experience and conceptual knowledge. To address this issue, contemporary research often acknowledges different subtypes of affordances. Some of these subtypes relate to specific kinds of actions, such as the use of the term *micro-affordance* to describe the potential for a grasping gesture resulting from perceiving an object's size, shape, location, and orientation (e.g., Ellis & Tucker, 2000). Other subtypes are distinguished by the origin and nature of the information used to establish the affordance. For example, an *episodic affordance* has been used to describe a potentiation for action that is not stable, but which changes as a function of an object's relative location in space (e.g., determining whether left-handed vs. right-handed manipulation would be

more likely for a human agent to attempt, see Glenberg, Becker, Klötzer, Kolano, Müller & Rinck, 2009).

A third question-- following on the heels of the idea that there are different subtypes of affordances-- is how different streams of information are prioritized or combined when a perceiver apprehends potential actions for objects. Consistent with the familiar notion of priming, research has suggest that particular affordances can come to be more prominent than others simply as a result of their earlier consideration. This can be shown in persistence effects whereby the classification of objects according to one affordance slows their subsequent classification using another affordance (Ye, Cardwell & Mark, 2009). Another line of work explores possible *a priori* differences in the priority or status of certain subtypes of affordances. For example, one proposal is that more primitive (nonconceptual) affordances such as the assessment of graspability may be computed more quickly that affordances relying on stored knowledge (e.g, the characteristic function of a familiar tool-- see, e.g., Vingerhoets, Vandamme, & Vercammen, 2009). The relationship between certain kinds of affordances and conceptual information has also been explored in neuropsychological work. For example, studies of perceptual deficits have illustrated cases where an individual's assessed knowledge of the appropriate actions for an object is relatively preserved despite an inability to correctly name the object, as well as the opposite pattern (Riddoch & Humphreys, 1987; Riddoch, Humphreys, & Price, 1989). Interestingly, however, a judicious examination of these patterns does not seem to warrant the conclusion that informational streams operate fully autonomously, but instead are linked via rather complex forms of interaction (see Humphreys, 2001).

Although the three themes mentioned above represent only a sampling of how affordances are explored in research on perception and action, they provide useful starting points

for considering the connection between affordances and language behaviour. This is due to their relationship to central questions in studies of real-time sentence comprehension, namely the extent to which so-called core aspects of processing proceed with or without information from the specific situational context, the notion that processing is subserved by multiple information sources, and questions regarding the time-course and architecture underlying the combination of different informational streams. These themes will be evident at various points in the review of studies that follows.

2. Affordances in visually situated language comprehension.

Visually situated language use differs from language occurring in other contexts in that it is used to denote entities, events, and states in the physical here-and-now. In this circumstance, various kinds of information are available that are either unavailable or are less available in instances where language is not situated. This includes details about the spatial position, size, and orientation of physical objects, as well as other outwardly-visible characteristics, all of which can be relevant for gauging the potential for certain forms of action. Given the availability of this information, it seems reasonable to think that affordances may have a distinct and important role to play in these contexts. What might this role be? One (perhaps obvious) possibility is a reduction in the amount or kind of information that would otherwise be explicitly stated using language. However, an effect of this type is probably not specific to affordances in particular, but would apply generally to any kind of information acquired via vision (e.g., the color of objects) or other sensory modalities. A more specific consequence that is explored in much of the research described below is that visually derived potentiations for action can

somehow help structure the nonlinguistic context, or *domain of interpretation* that sets bounds on interpretation of linguistic expressions.

Before venturing into this discussion, I will note that it is important to be mindful about core differences between perception and language, and to consider these carefully when reflecting on how affordances might be "plugged in" to language processes. For example, language is extremely nubile when it comes to the matter of perspective. Events can be expressed in the first, second, or third person, via direct or indirect quotation, and in ways that highlight either the endpoint or the starting point of actions (e.g., *Helmut gave Zelda a letter* vs. *Zelda received a letter from Helmut*). Affordances, on the other hand, are by definition a *first person* phenomenon: our apprehension of objects, participants, and potential actions within a visual scene is not the same as for other perceivers, and this apprehension will reflect the particular perspective delivered to the brain by our sensory systems. As a result, it is not unreasonable to expect that genuine affordances (as construed within work on perception) would be relevant to only selected aspects of situated language processing. Although cases involving second- or third-person effects (e.g., a sensitivity to the affordances apprehended by *another* individual) are interesting to consider, they cannot stem from precisely the same sensory-perceptual origins and instead will depend on higher-order representational systems that can yield a "simulated" perspective.

i. Mapping referring expressions to real-world referents

Referential expressions such as *the dog* or *the fork on the left* are interpreted against a circumscribed context (domain of interpretation) that sets limits on the nature and range of referential candidates for an expression as well as the contrasting alternatives that motivate the speaker's use of a particular description. In visually situated contexts, the set of perceptible

entities provides one obvious source of information for defining the domain of interpretation. Research has shown, however, that this is only a starting point, and the linguistically relevant domain for the semantics of a given expression is typically *narrower* than what is provided by the objects that have been visually indexed within a given scene. For example, there is now considerable evidence that information about actions and events that is encountered elsewhere in an utterance provides important constraints on the extent of the domain used in the interpretation of referring expressions. To illustrate, Altmann and Kamide (1999) examined how verbs in sentences such as *The boy will eat the cake* can influence referential interpretation in advance of hearing the noun. Listeners heard these types of sentences while viewing schematic scenes composed of clip-art images (see the chapters by Spivey & Huette and Pykkönnen & Crocker for additional discussion of the general methodological paradigm). The pattern of eye movements showed that listeners could use verb information to anticipate the referent in situations where only one scene object was compatible with the verb-evoked action (e.g., in a scene with only one edible object, given the verb *eat*). This suggests that event information in an unfolding utterance is used incrementally to (re-) define the contextual domain of interpretation, a concept that evokes a general theory of contextual "bookkeeping" advocated earlier by theorists such as Ballmer (1981):

> The general setting for bookkeeping is the following: linguistic expressions to be interpreted are analyzed from left to right. Single words or morphemes […] operate as linguistic instruments on the contextual aspects or parameters. These contextual aspects are conceived as sets [...] There are various explicit or implicit changes which may be induced by the interpretation of the simple linguistic expressions (morphemes, words, maybe phrases) on the context parameters. (p. 39)

However, it would be premature to conclude that the effect of event information in this case necessarily reflects the influence of affordances. For example, one alternative explanation

involves the notion of *selectional restrictions*— linguistically encoded well-formedness constraints on the combination of sentence elements (in this case, certain verb-noun pairings). A similar explanation is that the effect could be driven by stored and abstracted conceptual knowledge of the co-occurrences between actions and certain kinds of entities. For both of these explanations, the nature and source of the relevant information differs from a genuine affordance in that it abstracted and mentally stored. In other words, there is no evidence that comprehenders were computing the kind of "situation-specific" and "complementary" action relation between agents and the environment that characterizes an affordance as the term is understood within ecological perception. Another important point here is that the listener's own capacity for the denoted action (e.g., eating the cake) is not relevant. Rather, the listener is interpreting the described action in relation to a third-person participant mentioned in the discourse. This also limits the ability to understand the effect in terms of affordances.

A study by Kako and Trueswell (2000) using a more situated and first person experimental methodology provides an interesting analogue to Altmann and Kamide (1999). In this study, participants followed spoken instructions to execute an action involving one of several real objects located on a table top (e.g., *Now I want you to fold/pick up the towel*). As in Altmann and Kamide, the verb was manipulated such that it might be compatible with only the intended referent (e.g., *fold*) or all the objects present (e.g., *pick up*). The results were largely congruent with those from Altmann and Kamide's storyboard-like look-and-listen technique. At the verb, listeners rapidly narrowed consideration to action-compatible referents, allowing anticipatory identification of the intended target when only this object was compatible with the action. However, although Kako and Trueswell's experimental scenario does involve actual actions that are evaluated and executed by the perceiver/listener, it is still possible that mentally

stored information (selectional restrictions or conceptual associations) serves as the principal driver of the effect. To test whether affordances have an independent influence on referential interpretation in visual contexts, it seems necessary to examine cases involving *idiosyncratic* physical properties that are not among the properties that seem likely to be stored as part of the abstract conceptual representation for an object category.

In a study similar to Kako and Trueswell (2000), Chambers, Tanenhaus, Eberhard, Carlson and Filip (2002) used instructions whose predicate terms were of the type *put X inside Y.* Unlike lexically rich terms such as *eat* or *fold*, the lexical constraints stemming from the predicate information are comparatively uninformative in terms of co-occurrence associations with particular noun phrase referents. However, similar to the effect found with more lexically rich terms like *eat* or *fold*, interpretation at the point of the preposition *inside* showed rapidly-defined expectations for referents with action-relevant properties (namely open containers). Further, a clear influence of idiosyncratic physical affordances was found when the object array contained multiple containers matching the target name. Figure 1 depicts an example trial from the relevant experiment in which the corresponding instruction was *Pick up the cube. Now put it inside the can.* If we assume that the linguistically relevant domain is defined using the visually-present object array and the predicate information in the instruction (limiting the domain to those objects that are capable of containment), the second part of the instruction should be formally ambiguous due to the lack of a unique referent for the expression *the can*. However, one of the experimental manipulations varied the size of the theme object (the cube) such that it could or could not fit inside the smaller member of the target pair. When a small version of the theme object was used, listeners' eye movements and behavioral responses reflected the apparent ambiguity such that consideration of both cans was evident (and listeners were unsure what to

do). In contrast, when a large theme object was used (entailing that the smaller container no longer afforded the action evoked by the verb) the smaller alternative was excluded from the referential domain of interpretation, allowing the instruction to became functionally unambiguous. In this case, listeners' eye movements and actions showed no difficulty in selecting the intended target referent.

An additional manipulation involved using *indefinite* noun phrases in the second part of the instruction (e.g., *Pick up the cube. Now put it inside a can*). This was included to more firmly establish that the restricted domain reflected in the participants' actions is truly reflective of the mental representation used in the semantic evaluation of linguistic expressions, and does not simply reflect a type of task-based strategic response. Importantly, the felicity conditions resulting from the multi-referent context and the imperative *put_inside* instruction encourage a so-called "choice" interpretation for the indefinite noun phrase, such that it might be paraphrased as *one of the bowls*, which clearly presupposes the presence of multiple bowls. (Notice the interpretation here is distinct from indefinites in sentences such as *There is a strange man at the door*.) If the affordance-based exclusion of the smaller candidate referent genuinely reflects the linguistic domain used in defining the scope of the indefinite, listeners should experience confusion when the visual context contains the large version of the theme object. Indeed, this is what the data showed, suggesting that affordance-defined domains for action in this situation are in fact the same domains used in the semantic evaluation of linguistic elements.

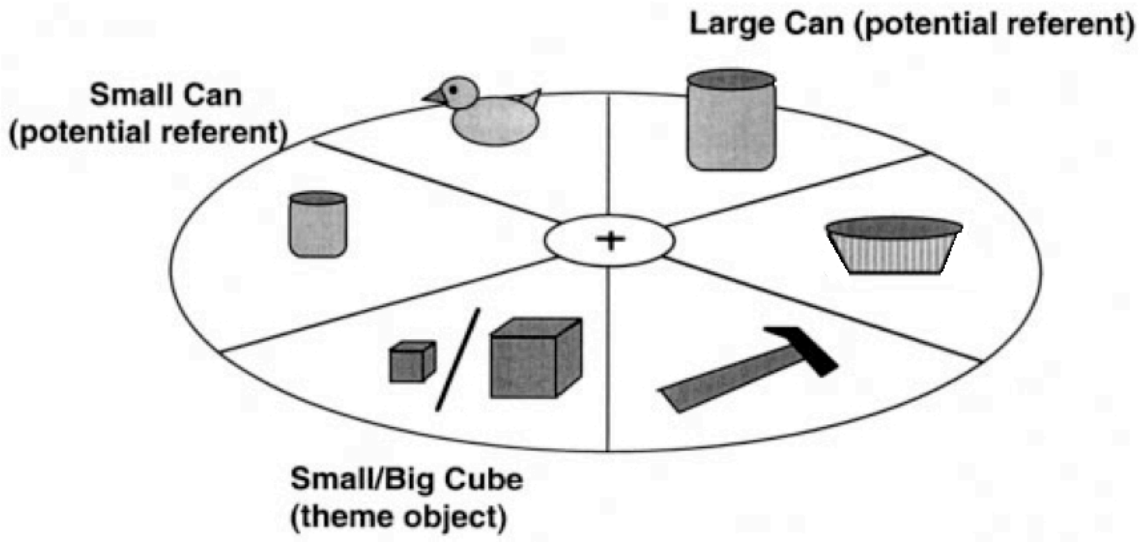

Figure 1: Example object array from Chambers et al. (2002), Experiment 2 (reprinted with permission).

Although not all instances of situated language use are likely to involve fully isomorphic domains-for-action and domains-for-language, the findings in this experiment provide useful starting points for additional exploration of this issue. The outcomes also highlight a clear back-and-forth interplay between linguistic and scene-based information sources to define the circumscribed attentional set. First, the visual scene provides a starting point by setting out the number and kinds of entities present. The predicate information, when heard, combines with this information to restrict consideration to container objects. Once this information (and the corresponding action) is known, another kind of scene-based information comes into play, namely the volumetric cues that allow an apprehension on the part of the listener as to which objects possess the relevant affordances for the evoked action. The interpretation of the subsequent linguistic expression is then guided by this information. It seems then that representations of the "context" and the "utterance" are shifting together in time in a kind of

lockstep relationship, where changes in one type of representation are spurred on by changes in the other. This situation illustrates a central tenet in Dynamical Systems approaches to cognition, namely that intelligent behavior is subserved by the continuous interaction of multiple tightly interconnected representations that *co-evolve* in time (e.g., van Gelder & Port, 1995).

ii. Recognizing spoken words

We have seen evidence that the mapping of linguistic expressions to nonlinguistic entities can be influenced by the perception of the affordances of these entities in relation to expressed actions. However, an important focus of research on real-time language comprehension concerns whether and how contextual factors penetrate into so-called core components of language understanding such as word recognition and computing the syntactic configuration of sentence elements. In classic frameworks, these processes have often been described as informationally encapsulated. On this account, the early moments of comprehension are free from the influence of contextual (nonlinguistic) information sources. Instead, these contextual constraints are integrated with the products of the initial processing phase only during a later time window.

When it comes to spoken word recognition, although the specific details vary across theoretical frameworks, most of these approaches assume that the recognition process is characterized by two features: the incremental uptake of information from the unfolding speech signal, and a competitive matching process in which information from the signal is dynamically evaluated against stored word forms in the mental lexicon. To illustrate, upon reaching the midpoint of the unfolding word *shark*, information in the signal provides a certain amount of evidence for a range of lexical alternatives including not only *shark* but *sharp, shard, spark,*

dark, etc. As each successive speech sound unfolds, the activation level of certain alternatives will be increased or reduced based on their cumulative match or mismatch with the available input.

One way to directly evaluate claims about informational encapsulation in word recognition is to test the potential for contextual constraints to limit the set of active lexical competitors as their incremental evaluation proceeds in time. Evidence from studies of spoken language in visually situated contexts has in fact provided evidence against a strong version of the encapsulation account. For example, event information from verbs encountered earlier can be used to reduce the consideration of phonetically similar word candidates as the signal unfolds in time (Dahan & Tanenhaus, 2004; Chambers & Cooke, 2009). However, predicate terms whose effects can be defined in terms of affordance-based information could provide an even more compelling case (due to the lack of other semantic associations between predicate terms and noun complements), and a clearer demonstration of cross-modal influences (i.e., the penetration of information from outside so-called language modules).

Revill, Tanenhaus, and Aslin (2008) tested this possibility in a study in which participants were first taught a novel lexicon that included words for complex geometric shapes as well as words for actions that could be performed on these shapes such as inducing movement (e.g., horizontal oscillation, clockwise rotation) and changing a shape's color/texture (e.g., grey-to-white, grey to speckled). The use of an artificial lexicon allowed the authors to stringently control for characteristics of the various words, including the number of alternatives whose sound pattern overlapped with a given target word (i.e., lexical neighborhoods), building on a methodology used by Magnuson, Tanenhaus, Aslin, and Dahan (2003). Notional affordances for the shapes were taught to participants during an initial training phase. For example, participants

might learn that shapes composed of straight lines could undergo movement changes but not texture changes. During the test phase of the experiment, listeners viewed displays on a computer screen that contained a selection of the geometric objects and various icons denoting different possible actions. For each display, they heard a recorded instruction consisting of an action word followed by an object word, and responded by clicking on the relevant action icon and then the denoted object. Eye movement data showed that the newly learned affordances influenced the extent to which a competitor (e.g., a straight-edged *bupa*) was visually considered as a target shape whose name had overlapping speech sounds (e.g., a rounded-edged *bupo*) was heard in the unfolding instruction. For example, if the required action could be afforded by the bupo, but not the bupa, consideration of the latter was significantly reduced as the word *bupo* unfolded in time, despite the overlapping "bu" sounds.

More subtle measures of the influence of affordance information on bottom-up aspects of word recognition were then obtained by exploring target fixations in situations where competitors were not present in the visual display. Previous eye tracking research has shown that the visual identification of targets with dense lexical neighborhoods (i.e., target words with many similar-sounding phonological neighbors) is slowed compared to those with sparse neighborhoods, even though these alternatives are not visually present (e.g., Magnuson, Dixon, Tanenhaus & Aslin, 2007). This reflects the implicit competition mechanisms referred to earlier: when the set of similar-sounding alternatives is larger, the activation level associated with each candidate is correspondingly lower in a roughly zero-sum manner, thereby delaying the point at which a target candidate eventually reaches threshold on the basis of the unfolding input. Of interest was whether this effect was modulated when some of the target word's phonological neighbors denoted objects that could not afford the action described by the action term. In other

words, can the affordance information evoked by the action term work to reduce the size of the implicit competitor set by limiting the bottom-up activation of certain lexical candidates? The results indicated that the information provided by the affordance constraints did indeed work to suppress competition stemming from non-displayed competitors, demonstrating the involvement of affordance-based considerations in core aspects of language processing.

iii. Computing syntactic relationships

Temporary linguistic indeterminacy is not limited to the unfolding of spoken words but is also an element of real-time comprehension at the syntactic level. As utterances are encountered in time, the grammatical relationships among entire words and phrases are often compatible with multiple structural analyses. To illustrate, the phrase underlined in the unfolding utterance *The baker poured the egg in the bowl...* may indicate the location where an egg is being poured, or may serve to indicate which of several eggs is intended. Information following the indeterminate phrase often disambiguates the intended grammatical relationship (e.g., …*while stirring continuously* vs. …*over the flour*). A substantial body of evidence has shown that listeners typically assign a provisional grammatical analysis to ambiguous phrases as soon as they are encountered, rather than waiting for disambiguating language. One of the core goals of research on syntactic-level comprehension is to characterize the kinds of information used to make these provisional decisions, and to understand how these information sources are integrated in real time as an utterance unfolds.

In many cases, comprehenders' initial understanding of an ambiguous phrase can be linked to the fact that the phrase follows a singular definite noun phrase (e.g., *the egg* in the preceding example). These expressions are typically used to refer to a uniquely identifiable

referent, and the amount of descriptive information necessary to achieve this goal will depend on the presence of other possible candidates in the referential context. Whether or not there is an expectation for additional descriptive information can influence the grammatical role initially assigned to an ambiguous phrase (see, e.g., Crain & Steedman, 1985). For example, if several eggs are present, comprehenders will initially interpret the *in the bowl* in *The baker poured the egg in the bowl…* as a modifying phrase because the simpler description *the egg* is not sufficient to specify which egg is intended. In contrast, if only a single egg is present, information of this sort would not be necessary and comprehenders would initially understand the phrase as indicating the intended location. But how does this "number of potential referents" phenomenon link up to the perception of affordances?

Recall that the evidence discussed earlier shows that listener/perceivers use lexical semantics in combination with visually defined affordances to restrict the visual scene to compatible referential candidates, and that these restricted domains are used to determine the potential uniqueness of entities in relation to a definite expression. It follows, then, that affordances may play a role in influencing expectations about whether additional information is required to achieve referential success as the components of a description are successively encountered. In one study, Chambers, Tanenhaus and Magnuson (2004) recorded eye movements as listeners followed instructions to interact with real objects in a visual display. Instructions were of the type *Pour the egg in the bowl over the flour*, where the first prepositional phrase (*in the bowl*) is temporarily ambiguous with respect to its syntactic role (location vs. modifier). The important manipulation in the visual display was whether both visually present candidates (e.g., an egg in a bowl and an egg in a glass, see Figure 2) could afford the described action (e.g., they were both in liquid form) or whether only one candidate possessed the relevant

affordance (i.e., the egg in the glass was still in the shell and hence unpourable). The critical measure was whether listeners initially misinterpreted the ambiguous phrase as specifying the intended destination, as measured by whether they fixated the empty bowl (the "false" destination for the object to be moved) upon hearing the ambiguous phrase in the unfolding instruction. The results showed that a destination interpretation was adopted when only one referential candidate afforded the described action. When both alternatives were compatible, a modifier interpretation was adopted, and fixations to the false destination were no more than what was observed when the instruction was linguistically unambiguous (e.g., *Pour the egg that's in the bowl over the flour.*).

Apart from illustrating the role of affordances in core aspects of syntactic processing, the outcomes help address a question that was left only partially answered in other work. Although affordances were shown to constrain the domain of interpretation for simple definite referring expressions (Chambers et al., 2002), one might argue that a modified description would have been more expected or effective (e.g., *Now put it inside <u>the large can</u>* would be a more felicitous instruction corresponding to Figure 1, even when it was clear that the smaller can in the display could not afford the denoted action). The results from the "pour the egg" study suggest that this is not the case. Specifically, listeners actually dispreferred a modifier interpretation for the *in the bowl* phrase when only one of the two lexically-compatible candidates was physically compatible with the stated action, and instead temporarily misinterpreted this phrase as specifying the intended destination. This outcome reinforces the idea that the circumscribed domain for planning and executing actions appears to be the same domain used in the semantic evaluation of linguistic expressions, at least in these types of situations.

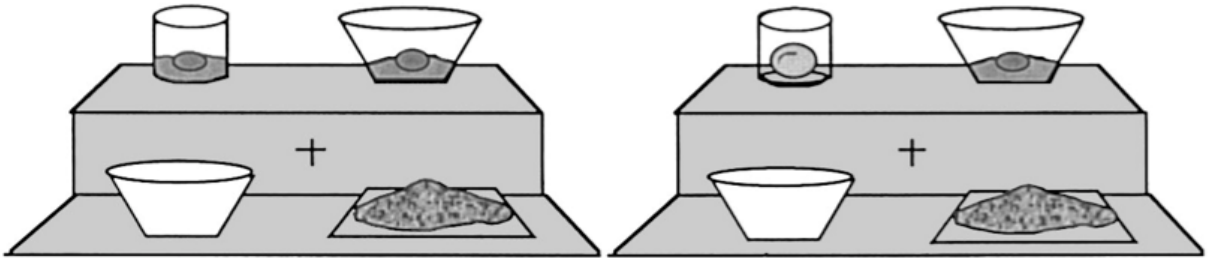

Figure 2: Example object array from Chambers et al. (2004), Experiment 1 (reprinted with permission).

The finding that the domain-restricting effect of affordances operates rapidly enough to influence both incremental spoken word recognition and real-time syntactic processing has clear implications for our understanding of the mental architecture underlying language comprehension processes. As mentioned before, results from earlier studies, as well as prominent theoretical frameworks, had championed the idea that the early moments of processing drew only on selected information sources intrinsic to the linguistic system. The immediate influence of informational constraints that are clearly nonlinguistic in nature (i.e., provided by a different sensory modality) and which reflect broader behavioral goals (i.e., the intention to execute a spoken instruction by means of physical action) provides a significant challenge to this view. These findings are instead consistent with the idea that the coordination of information during real-time language comprehension is opportunistic-- drawing on any relevant cues according to their availability-- and is characterized by highly parallel processing such that various information streams are integrated seamlessly as words and sentences unfold.

iv. Interpreting language in conversational contexts

To this point, the role of affordances in visually situated language comprehension has been illustrated in experiments where the assessment of potential actions is made by a lone perceiver operating within a physical environment containing a number of objects and little else. In these studies, language is provided via recorded speech or an experimenter who is not visible within the contextual environment and who is not interacting with the listener/perceiver in any real sense. With this in mind, it is important to recall that the traditional definition of an affordance involves a consideration of the bodily capacities of an organism in relation to a perceived environment, where "environment" is understood in a reasonably broad sense, and not just an array of manipulable objects. In the canonical situated setting for communication, the environment includes other individuals, namely other conversational participants. A consideration of this kind of context provides an opportunity to examine affordances from yet another angle. In particular, we can ask how the presence of another individual creates additional and possibly more complex kinds of potential actions whose detection nonetheless occurs via perception.

At perhaps the most coarse-grained level, one type of potential action created through the presence of an interlocutor is simply the basic act of beginning to speak, or, more specifically, assuming the role of speaker within the regime of conversational turn taking (see Greeno, 1994). During conversation, speakers produce certain perceptible linguistic and nonlinguistic behaviors that help to signal the end of their conversational turn such as pausing, slowing of speech rate, and a drop of pitch, as well as changes in gesture and gaze behavior. These physical cues can lead the perceiver (in the current role of listener) to calculate the point at which to effect a certain

kind of action, namely to take over as the active speaker. The management of conversational turn taking can therefore be argued to involve the detection of affordances at some level.

A topic that builds more directly on the findings discussed earlier concerns the perceiver/listener's apprehension of the actions available to *another* individual, and how this might influence aspects of language understanding. Such a process would obviously depend on certain representational transformations in the sense that it no longer involves detecting potentiations for action from a first-person perspective. In one study, Hanna and Tanenhaus (2004, see also the chapter by Brown-Schmidt, this volume, for related work) conducted an experiment varying the spatial accessibility of objects in relation to the potential actions of a speaker, and examined how this influences the mapping of definite referring expressions to these objects. In the experiment, participants assisted a confederate (playing the role of "cook") in the task of following a recipe. The materials for the recipe were spread across a physical workspace such that some in the participant's area, others were in the cook's area, and some were accessible to both. The cook provided the participant with a series of instructions, some of which were referentially ambiguous, e.g., *Could you put the cake mix next to the mixing bowl*?, spoken when two boxes of cake mix were physically present. In the critical test conditions, one of the two potential referents was reachable to the participant, and the other was reachable to the cook. Further, the instruction was spoken at a moment when the cook's hands were either empty or happened to be occupied with the task of holding a tray. This manipulation varied the presence of an action-based cue that could potentially restrict the domain of interpretation for the listener. Specifically, when the cook's hands were empty, a participant should infer that the intended referent is the one in the participant's own area, because it would be implausible for the cook to request an object that she could actually reach herself. However, when the cook's hands were

full, her capacity to reach either cake mix is impeded, and consequently both alternatives become a plausible candidate for the expression (making the expression referentially ambiguous).

Eye movement data, as well as participants' overt behaviors, suggested that the referential domain of interpretation was indeed quickly defined on the basis of these ascribed affordances, in combination with assumptions of cooperative communication. Thus, the actions we perceive as being available to other individuals are used to define linguistically relevant domains of interpretation. As mentioned earlier, however, this outcome is likely to be the product of comparatively more complicated forms of mental representations compared to the "first person" apprehension of affordances. This particular point is substantiated by developmental patterns in the time course of processing. Although the behavior and overall gaze patterns of 3- and 4-year-olds in the same kind of experimental task are like those observed in adults by Hanna and Tanenhaus (2004), children's eye movement patterns in the very earliest moments of processing did not reflect the same ability to rapidly restrict the domain of interpretation (Collins, Graham, & Chambers, 2012).

In the case of conversational interaction, the actions that are perceptible to a listener/perceiver also include actions produced by a speaker that involve the actual mechanics of speech production. If listener/perceivers routinely take into account the actions available to other individuals, it is therefore plausible that something perceived as impeding the speaker's articulatory processes could correspondingly affect the comprehender's interpretation of sounds being produced. In an intriguing study of perceptual learning, Kraljic, Samuel and Brennan (2008) examined the influence of this type of information. The authors' starting point was an established re-tuning effect that occurs when listeners accommodate to idiosyncratic speaker-based variation in the pronunciation of speech sounds. In one condition, participants viewed an

audiovisual recording of a speaker who tended to pronounce words in a way where the sound /s/ was more "sh"-like than the articulatory standard. With repeated exposures, this resulted in a boundary adjustment in the phonetic category such that the listener identified "sh"-shaded tokens as belonging to the /s/ category. Critically, however, this perceptual learning effect was blocked when visual information allowed the speaker's pronunciation patterns to be attributed to an incidental cause, namely a pen that the speaker happened to have in her mouth when uttering /s/-containing words. (The audio track in the videorecording was the same as for the "pen-free" condition, ensuring that any difference in learning patterns could not be attributed to differences in the acoustic signal.) This outcome provides yet another perspective on how perceived capacities for action can penetrate into core linguistic processing.

3. Are affordances really the right explanatory concept?

To this point I have not discussed another body of research that draws on the notion of affordances, namely the burgeoning literature exploring the *embodiment* of language processes. Most generally, this work adopts a perspective in which semantic-conceptual meanings are understood to be "grounded" in perception, and language comprehension is thought to involve a process of covert perceptual and/or motor simulation in response to described entities and events (see Zwaan & Kaschak, 2009 for an overview). Although this approach stands in contrast to frameworks assuming highly abstracted and amodal propositional structures, it reflects a strong degree of continuity with Mental Models accounts, which have frequently appealed to aspects of perception to characterize the mental representations generated from language (e.g., spatial location, attentional focus, foregrounding/backgrounding, perspective, among others: see Garnham & Oakhill, 1996; Johnson-Laird, 1983; Zwaan & Radvansky, 1998). The specific role

of affordances in the embodied approach is evident in studies examining combinatorial meaning in sentences, where the goal is to understand how comprehenders establish links between denoted actions or states and the individuals and objects involved in these eventualities. For example, Glenberg and colleagues (e.g., Glenberg & Robertson, 1999, 2000; Kaschak & Glenberg, 2000) have argued that mental simulations for the affordances of linguistically-evoked entities are used to "mesh" representations of these entities with described actions, and that this process provides a superior account to other accounts of how these semantic linkages are identified (e.g., those based on distributional co-occurrence statistics or abstract propositional representations).

Despite the thematic similarities, there are several ways in which this body of research reflects a focus that is distinct from the work reviewed so far in the current chapter and the general theme of this book. For example, the studies mentioned in the preceding sections reflect a position in which perceptual information and its corollaries function as an *accompanying stream of information* derived from the co-present visual context, rather than an outcome of language interpretation (i.e., a perceptually-rich internal simulacrum of a described state or event). This is not to say, however, that these studies described in this chapter endorse the idea that mental representations of language are disembodied in the sense of being abstract propositional structures, etc. It is possible, for instance, that the rapid integration of linguistic and visual information is subserved in part by shared systems and/or representational formats (see, e.g., Spivey & Richardson, 2009). Another difference is that work on the embodiment of language tends to be more concerned with the more final representational instantiations of sentences, rather than the on-line processes that handle temporally evolving (and hence incomplete/indeterminate) linguistic input. This distinction may be best appreciated as a matter

of degree rather than an absolute, but it is nonetheless important in understanding the goals that motivate the respective approaches to understanding language comprehension mechanisms.

What embodied approaches can directly contribute to the current question is to encourage us to more stringently consider what it means if affordances can also influence language interpretation in a simulated form, in the absence of concurrent visual processing. One (perhaps too easy?) conclusion is that studies of embodied language and visually situated language together yield a unified and consistent empirical argument highlighting the importance of affordances for comprehension processes. Another response is to consider instead the *differences* across these kinds of studies, and whether these differences point to a common denominator that is in fact less tied to aspects of sensory-perceptual information. We have already broached the question, for instance, of whether the direct apprehension of affordances in the "cognition-free" manner suggested by Gibson is in fact possible, and it seems clear that the mental simulation of affordances during discourse comprehension will also necessarily involve higher-order mechanisms. The studies discussed in the preceding sections also reveal ways in which higher cognition is involved in the perception of action-relevant properties (even in visually-situated contexts). This was evident in studies where affordances are calculated in relation to the speaker's actions (Hanna & Tanenhaus, 2004), or to third person clip-art characters (Altmann & Kamide, 1999, and others). Yet another example comes from work showing that two-dimensional clip-art images of objects can evoke information about the typical mode of manipulating objects. For example, an image of a piano evokes the manual behaviors involved in using the real-world analogue of that object, and in turn activates the concept *typewriter* due to the common mode of bodily interaction (Myung, Blumstein & Sedivy, 2006; see also Yee, Huffstetler, & Thompson-Schill, 2011). This phenomenon neatly bridges the kinds of effects

observed in studies using visually-present real objects and those found in work studying linguistically-driven mental simulations of affordance information.

If, then, affordances in language comprehension are (in many cases) tinged by influences of higher-level cognition, it may make sense to consider whether there is anything special about these affordances *per se*, or whether they simply provide a particularly accessible demonstration of the idea that real-time language understanding is ultimately an act of recognizing the communicative intent of an individual, and that linguistic information is supplemented by detailed forms of nonlinguistic information in achieving this goal. One way to address this question is to evaluate the primacy of sensory-perceptual information in relation to other information sources in the interpretation of natural language. This typically requires the use of linguistic stimuli containing more semantically- or pragmatically-nuanced lexical items or constructions that can reveal the potential for visually based information to be modulated or suppressed.

One study of visually situated language that provides some relevant data is an exploration of linguistic contrast by Sedivy, Tanenhaus, Chambers, and Carlson (1999). As background, the claim that language is "modal" and involves rich and detailed perceptual simulations would seem to commit us to the idea that representations generated on the basis of language are perceptually specific as opposed to abstract. It is thus reasonable to think that the internal representations for linguistic descriptions such as *heavy book, dog, new shoes*, etc. would involve some kind of well-defined visual exemplar. On this assumption, it would follow that there should be some detectable penalty when the referent of a linguistic description is not an ideal fit with this perceptual representation. Such a penalty was indeed detected in the Sedivy et al. study. Specifically, upon hearing a sentence such as *Pick up the tall glass*, listeners were slower to

fixate the target object in the display when it was a fairly normal-sized glass, compared to one that was perceptibly taller than an average glass (and would therefore be a better referent for the description *tall glass*).

However, an important finding in the study involved a situation in which the target glass was accompanied by a smaller contrasting glass in the visual scene. In this case, the specific size of the target referent glass no longer seemed to matter: listeners were as fast to identify it regardless of whether it was/was not an average glass or tall glass relative to the category norm. This reflects the pragmatic conditions of use for dimensional modifiers like *tall*: speakers tend to include them in referring expressions to differentiate objects from one another in a *relative* way rather than to simply ascribe some stable property to an entity. At a minimum, this suggests that the semantics and pragmatics of natural language can mute the importance of certain kinds of perceptual information, such as stored visual standards for object categories.

Other work has used the phenomenon of referential anticipation to explore the extent to which perceptual information plays a dominant role in referential processing. Chambers and San Juan (2009) investigated the interpretation of the transitive verb *return* in instructions such as *Now return the square to area 3*, occurring within a sequence of several instructions. Like the verb *move, return* expresses an overt physical action involving concrete objects, making it relevant for exploring the topic of perceptually grounded aspects of language interpretation. However, it is also clearly a presuppositional term, whose semantics requires a particular background condition to have been satisfied (namely the previous displacement of the denoted object). These types of expressions provide an opportunity to explore the influence of other kinds of constraints alongside perceptual and action-defined information.

The study revealed that, even in the earliest moments of comprehending the verb, the consideration of referential candidates was influenced by various nonperceptual factors including the inferred purpose behind the original object displacement, and whether an earlier displacement was considered to be relevant to communicative goals. For instance, when a participant had to move a certain display object "incidentally" to enable an object of interest to continue along a particular path, the incidentally-moved object was not considered when listeners heard the verb in a subsequent *Now return the…* instruction, even though this object unambiguously satisfied the perceptual-level affordances for the verb *return* (i.e., it was known/perceived to be previously displaced just seconds before). The perceptually defined affordance of "being returnable" was apparently muted for this object by higher-level factors pertaining to goal relevance.

These results, as well as findings from some other visually situated studies (see, e.g., Altmann & Kamide, 2009; Wolter, Skovbroten Gorman & Tanenhaus, 2011), indicate that the use of perceptually-derived information can be readily suppressed in reaction to certain semantic, pragmatic, and discourse-based requirements during language interpretation. At the same time however, other work has illustrated situations where perceptually derived information trumps other potentially relevant knowledge for processes such as the linking of actions to entities (e.g., overruling stored stereotypic associations between specific actions and event participants, see Knoeferle & Crocker, 2007). Taken together, these studies highlight the need for a coherent middle ground that more readily acknowledges the fact that the same information can have different effects in different circumstances. In some cases, this approach will require more attention to the diversity of meanings expressed in natural language, (particularly those that do not relate to aspects of perception or action). On other cases, this will likely require careful

thinking about the nature of experimental tasks and their relationship with the range of goals and behaviors found in real-world communicative contexts.

4. Summary

The notion of affordances is explicitly and implicitly present in a broad range of experimental studies of situated language comprehension. This concept has informed research on both specific and general topics, including modularity in core linguistic processing, the question of how linguistically-relevant context is defined (and re-defined over time), the real-time integration of qualitatively different types of information, the linking of referential entities to predicate terms, and the content of mental representations for language. It is unclear at this point, however, whether affordances-- and other perceptually-derived information-- play a particularly distinct and privileged role in language processes (even in visually situated situations), or whether they just provide a particularly salient illustration of the fluid and contextually-sensitive character of the human capacity to interpret language in real time. What is clear, in contrast, is that the recent interest in affordances marks an important shift towards appreciating idiosyncratic features of objects and events in aspects of language understanding. The fact that this interest accompanies the increasing use of methodologies for studying visually situated spoken language is not surprising. In reading paradigms, it is difficult to provide detailed information about the nature of denoted entities and actions, in part because of the obvious need to provide more text (thereby increasing the length of each trial and consequently limiting the number of observations and/or conditions) and also because the explicit provision of this information may lead to unwanted inferences about its importance or relevance. Standard theoretical models for phenomena such as linguistic reference have also been somewhat

restricted when it comes to the situation-specific features of actions, states and entities. For example, the framework provided by representational models such as Discourse Representation Theory (Kamp & Reyle, 1993) tends to adopt a comparatively atomic representation of referents, rather than a more molecular view in which these referents are represented along with their various attributes, including idiosyncratic properties. Visually based psycholinguistic paradigms, in contrast, demonstrate how this information comes to be incorporated into mental representations without effort or fanfare, as a basic by-product of perceiving the broader contextual environment in which language occurs.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Altmann, G. T. M., & Kamide, Y. (2009). Discourse mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition, 111*, 55-71.

Ballmer, T. (1981). Context change and its consequences for a theory of natural language. In H. Parret, M. Sbisà, & J. Verschueren (Eds.), *Studies in language companion series: Vol. 7. Possibilities and Limitations of Pragmatics* (pp. 17-55). Amsterdam: John Benjamins.

Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: Sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory & Cognition, 35*, 1029-1040.

Chambers, C. G., & San Juan, V. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. *Cognition, 108*, 26-50.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language, 47*, 30-49.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*, 687-696.

Collins, S.J., Graham, S.A., & Chambers, C.G. (2012). Preschoolers' sensitivity to speaker action constraints to refer referential intent. *Journal of Experimental Child Psychology, 112,* 389-402.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives* (pp. 320-358). Cambridge, UK: Cambridge University Press.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping form sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*, 498-513.

Ellis, R., & Tucker, M. (2000). Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology, 91,* 451-471.

Garnham, A., & Oakhill, J.V. (1996). The mental models theory of language comprehension. In B.K. Britton & A.C. Graesser (Eds.), *Models of Understanding Text* (pp. 313-339). Hillsdale, NJ: Erlbaum

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67-82). Hillsdale, NJ: Erlbaum.

Gibson, J. J. (1979). *The ecological approach to visual perception*. New York: Houghton Mifflin.

Glenberg, A. M., Becker, R., Klötzer, S., Kolano, L., Müller, S., & Rinck, M., (2009). Episodic affordances contribute to language comprehension. *Language and Cognition, 1* , 113-135.

Glenberg, A. M., & Robertson, D. A. (1999). Indexical understanding of instructions. *Discourse Processes, 28*, 1–26.

Glenberg, A. M. Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language 43*, 379–401.

Greeno, J.G. (1994). Gibson's affordances. *Psychological Review, 101*, 336-342.

Hanna, J.E. & Tanenhaus, M.K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science, 28*, 105-115.

Humphreys, G. (2001). Objects, affordances, action! *The Psychologist, 14*, 408-412.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, UK: Cambridge University Press.

Jones, K.S. (2003). What is an affordance? *Ecological Psychology, 15,* 107-114.

Kako, E. & Trueswell, J. C. (2000). Verb meanings, object affordances, and the incremental restriction of reference. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 256-261). Hillsdale, NJ: Erlbaum.

Kamp, H. & Reyle, U. (1993). *From discourse to logic*. Dordrecht, NL: Kluwer.

Kaschak, M.P., & Glenberg, A.M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language, 43*, 508-529.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language, 57*, 519–543.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science, 19*, 332-338.

Magnuson, J. S., Dixon, J., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 133-156.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (2003). The time course of spoken word recognition and learning: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*, 202-227.

Myung, J.-J., Blumstein, S.E., & Sedivy, J.C. (2006). Playing on the typewriter, typing on the piano: manipulation knowledge of objects. *Cognition, 98*, 223-243.

Revill, K. P., Tanenhaus, M. K., & Aslin, R. N. (2008). Context and spoken word recognition in a novel lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1207-1223.

Riddoch, M. J., & Humphreys, G. W. (1987). Visual object processing in a case of optic aphasia: A case of semantic access agnosia. *Cognitive Neuropsychology, 4,* 131-185.

Riddoch, M. J., Humphreys, G. W., & Price, C. J. (1989). Routes to action: Evidence from apraxia. *Cognitive Neuropsychology, 6,* 437-454.

Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., & Carlson, G.N. (1999). Achieving incremental interpretation through contextual representation. *Cognition, 71*, 109-147.

Shaw, R.& Turvey, M. T. (1981). Coalitions as models for ecosystems: A realist perspective on perceptual organization. In M. Kubovy & J. Pomerantz, *Perceptual organization* (pp. 343-415). Hillsdale, NJ: Erlbaum..

Spivey, M. & Richardson, D. (2009). Language embedded in the environment. In P. Robbins and M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 382-400). Cambridge, UK: Cambridge University Press.

van Gelder, T., & Port, R. (Eds.), (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

van Leeuwen, L., Smitsman, A., & van Leeuwen, C. (1994). Affordances, perceptual complexity, and the development of tool use. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 174–191.

Vingerhoets, G., Vandamme, K., & Vercammen, A. (2009). Conceptual and physical object qualities contribute differently to motor affordances. *Brain and Cognition, 69*, 481-489.

Wolter, L., Skovbroten Gorman, K., & Tanenhaus, M. K. (2011). Scalar reference, contrast and discourse: Separating effects of linguistic discourse from availability of the referent. *Journal of Memory and Language, 65,* 299-317.

Ye, L., Cardwell, W., & Mark, L. (2009). Perceiving multiple affordances for objects. *Ecological Psychology, 21,* 185-217.

Yee, E., Huffstetler, S., Thompson-Schill, S.L. (2011). Function follows form: Activation of shape and function features during object identification. *Journal of Experimental Psychology: General, 140*, 348-363.

Zwaan, R. A., & Kaschak, M. P. (2009). Language in the brain, body, and world. In P. Robbins and M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 368-381). Cambridge, UK: Cambridge University Press.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.