Available online at www.sciencedirect.com
ScienceDirect
Cognition 106 (2008) 1382–1407
COGNITION
www.elsevier.com/locate/COGNIT

Segmenting dynamic human action via statistical structure

Dare Baldwin [a,*], Annika Andersson [a], Jenny Saffran [b], Meredith Meyer [a]

[a] *Department of Psychology, 1227 University of Oregon, Eugene, OR 97403-1227, USA*
[b] *Department of Psychology, University of Wisconsin – Madison, 1202 W. Johnson Street, Madison, WI 53706, USA*

Received 5 January 2007; revised 27 June 2007; accepted 15 July 2007

Abstract

Human social, cognitive, and linguistic functioning depends on skills for rapidly processing action. Identifying distinct acts within the dynamic motion flow is one basic component of action processing; for example, skill at segmenting action is foundational to action categorization, verb learning, and comprehension of novel action sequences. Yet little is currently known about mechanisms that may subserve action segmentation. The present research documents that adults can register statistical regularities providing clues to action segmentation. This finding provides new evidence that structural knowledge gained by mechanisms such as statistical learning can play a role in action segmentation, and highlights a striking parallel between processing of action and processing in other domains, such as language.

Keywords: Action processing; Action segmentation; Statistical learning

* Corresponding author. Tel.: +1 541 346 4964; fax: +1 541 346 4911.
E-mail address: baldwin@darkwing.uoregon.edu (D. Baldwin).

0010-0277/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2007.07.005

1. Introduction

We are an intensely social species: social engagement permeates our daily lives, and interactions with others pervasively shape our physical and emotional well-being. Successful social functioning in everyday life depends crucially on skills for quickly and accurately analyzing the actions others are undertaking. Countless times in any given day we must make rapid judgments about others' current and likely future actions in order to engage with them effectively. Rapid, skilled processing of action[1] is also foundational to our general cognitive and linguistic functioning. In particular, our analysis of others' actions enables us to gain information about the world more generally – for example, information such as the safety, desirability, and functional properties of objects. As well, skill at action processing plays a crucial role in our ability to formulate linguistic descriptions of events we witness, as well as in building representations of the descriptions that others provide for us. It is obvious that cognitive and perceptual mechanisms of substantial complexity make skilled action processing possible (e.g., Blakemore & Decety, 2001; Frith & Frith, 1999), yet much remains to be learned about the nature of such mechanisms.

At the surface level, action is complex: Our bodies travel rapidly among a myriad of objects that we manipulate in diverse and often novel ways. One fundamental problem observers must solve in processing action is segmentation. In everyday action, our behavior tends to flow continuously, with few pauses to mark meaningful boundaries between distinct acts (Asch, 1952; Heider, 1958; Newtson & Enquist, 1976). Identifying distinct acts within the dynamic flow of motion is a basic requirement for engaging in further appropriate processing of the behavior stream. Among other things, extracting action segments is necessary to identify the kinds of action being undertaken (e.g., grasp, push, pull), to register novel combinations of known actions, and to learn words that label actions.

Prior research clarifies that adults readily segment continuous intentional action. If asked to indicate meaningful breakpoints in continuous, everyday action, adults agree about where boundaries separating distinct actions lie (e.g., Newtson, 1973), they can make such judgments at multiple levels within a hierarchy (e.g., Hard, Lozano, & Tversky, under review; Tversky, Zacks, & Martin Hard, in press; Zacks et al., 2001; Zacks & Tversky, 2001), and their judgments about such boundaries tend to coincide with their analysis of the intentions that actors are carrying out (e.g., Baird & Baldwin, 2001; Hard, Tversky, & Lang, in press; Zacks, 2004). For example, in the case of everyday actions such as kitchen clean-up, observers readily identify relevant segments at a fairly fine-grained level (including acts such as grasping a dish, grasping a faucet handle, and twisting the faucet handle), as well as at higher levels (e.g., washing a dish, hanging a towel), linked in a hierarchy to the smaller-action segments. Boundaries between actions emerge as psychologically relevant in both recall and on-line processing of dynamic action, and segmentation occurs spontaneously,

[1] In using the term "action," we are referring to the physical activity that people engage in as a route to fulfilling intentions and goals. We use the term "act" as a variant of "action."

even when not in any way necessary to the task at hand (e.g., Loucks & Baldwin, 2006). Even infants display some basic skill at segmenting dynamic human action (e.g., Baldwin, Baird, Saylor, & Clark, 2001; Saylor, Baldwin, Baird, & LaBounty, 2007). Finally, specific neurophysiological sites have recently been identified as active in segmentation of dynamic human action (Zacks et al., 2001; Zacks, Swallow, Vettel, & McAvoy, 2006). Together, these existing findings indicate that skill at detecting action segments plays a key role in processing of dynamic human activity. At the same time, as yet the available findings have provided little insight into the specifics of how observers of dynamic action identify relevant action segments within a continuous behavior stream. That is, the mechanisms enabling adults to extract segments from a continuous flow of activity have not been known.

There is reason to believe that a variety of processes operate in adults' action segmentation (e.g., Baird & Baldwin, 2001; Baldwin & Baird, 2001; Hard et al., in press; Newtson, Enquist, & Bois, 1977; Zacks, 2004). High-level "top-down" knowledge about the kinds of goals, intentions, and associated actions that are likely in given contexts should facilitate segmentation. For example, our prior knowledge about the kinds of goals and intentions often acted upon in kitchens leads us to expect segments such as dish-washing and returning items to cupboards and refrigerator. Such expectations, which include causal knowledge of motions requisite to satisfy intentions and goals, should assist us in identifying where such actions begin and end within the behavior stream.

Other kinds of knowledge that are more structural in kind might work in concert with high-level knowledge of intentions and goals to assist adults in segmenting continuous human action. The pursuit of everyday goals seems to introduce structured patterns into the flow of motion that actors produce (e.g., Newtson et al., 1977). If observers can detect structural regularities within motion that happen to correlate with the initiation/completion of individual intentional acts, this could assist them in segmenting the behavior stream independent of any high-level knowledge about the intention/goal content of the activity underway. The present research investigates one kind of structural knowledge – knowledge of sequential probabilities[2] – that people may be able to exploit to assist in discovering segments within dynamic human activity. Although we think it likely that intention/goal knowledge and structural knowledge work in concert to achieve segmentation in normal processing of everyday action, gaining definitive evidence for structural knowledge requires that any contribution from intention/goal content be eliminated as a source of information about the action segments at issue. This is thus the strategy we pursued in the present research.

1.1. Statistical learning could facilitate action segmentation

Statistical regularities could potentially be of assistance for identifying segmental structure within dynamic activity because some small-scale acts (e.g., grasp knife,

[2] In using the term "sequential probability," we are referring to the probability that Y follows X, computed by dividing the frequency of XY by the frequency of X. This provides a measure of how tightly linked X and Y are.

slice with knife) within the stream of behavior co-occur more frequently than others. Such sequential probabilities likely arise, at least in many cases, because the small-scale acts involved are causally linked in achieving a goal (e.g., in preparing stew, the motion of slicing a vegetable is frequently preceded by the motion of grasping a knife, whereas slicing a vegetable is only infrequently preceded by grasping a refrigerator door). It is likely, of course, that knowledge of the intentions, goals, and relevant causal motions on the observer's part would help to "bind" some adjacent actions together as a unit in the observer's processing. For example, knowing that (a) knives are useful for chopping carrots, (b) chopping speeds cooking, and (c) knives have little causal relevance to refrigerators, could help to make knife grasping and carrot chopping cohere, while refrigerator grasping and carrot chopping do not. However, even in the absence of relevant causal/intentional knowledge, sensitivity to the concomitant statistical regularities themselves might enable observers to group small-scale segments into relevant higher-level action bundles. Put another way, from the observer's point of view, a history of low sequential probabilities for two adjacent small-scale acts is a potential clue to segmentation at a higher level; that is, low sequential probabilities predict boundaries between distinct higher-level actions.

These ideas initially arose on analogy with recent findings in the language domain. Segmenting language, like action, requires discovering segments within a complex, dynamic stimulus stream. Human infants as well as adults can detect "word"-level segmental structure within a novel stream of syllables (small-scale segments that are familiar and readily extractable) via sensitivity to sequential probabilities across syllables (e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). Perhaps adults can similarly exploit statistical structure to discover higher-level segmental structural within a novel stream of dynamic activity.

Evidence for statistical learning in the visual domain lends plausibility to this hypothesis, given the strongly visual nature of action processing. A sizable body of research now indicates that infants as well as adults readily learn to detect predictable combinations (either temporal or spatial) of visual elements in arbitrary, novel displays (e.g., Fiser & Aslin, 2002a, 2002b; Hunt & Aslin, 2001; Turk-Browne, Jungé, & Scholl, 2005). On the other hand, however, it is difficult to generalize directly from this research to processing of human action. For one, detecting statistical regularities among simple visual shapes is very different from discovering regularities in extended displays of human activity. As well, considerable research suggests mechanisms for processing human action may be specialized in at least some respects (e.g., Shiffrar & Freyd, 1990); thus caution is warranted in generalizing from existing research on visual statistical learning to the domain of human action. A different reason for caution concerns the nature of the task used in many studies. Swallow and Zacks (submitted for publication) have made the point that paradigms such as the serial reaction-time task (e.g., Cohen, Ivry, & Keele, 1990; Nissen & Bullemer, 1987) – used in many demonstrations of visual statistical learning – involve learning linkages between predictably recurring visual elements and specific motor responses, with no way of determining the extent to which findings specifically reflect learning about regularities in the visual displays. Hence findings from the serial reac-

tion-time task are not necessarily directly informative regarding adults' ability to exploit statistical structure to segment the stimulus stream, which is the question of interest in the present research. Of course, studies of visual statistical learning were undertaken with very different aims than the present research, making it unsurprising that they do not clarify whether statistical learning skills might support action segmentation. In sum, the present research extends existing findings regarding visual statistical learning to examine whether adults can bring statistical learning skills to bear on dynamic action in ways that would support action segmentation.

Pioneering research by Avrahami and Kareev (1994), motivated by an insight very similar to the present hypothesis, documented that adults can learn to identify units within a semi-arbitrary sequence of events (e.g., randomly ordered clips from *Roadrunner* and *Coyote* cartoons) based solely on patterns of co-occurrence in their prior experience with a continuous sequence of those clips. These seminal findings further increase the plausibility of our hypothesis. At the same time, however, there are a number of reasons for questioning whether the Avrahami and Kareev research resolves the question of whether adults are sensitive to segmentation-relevant statistical structure in human action. First, the Avrahami and Kareev research employed highly artificial stimuli; for instance, several studies utilized cartoon animations. A third study employed a black-and-white studio film (a Jacques Tati holiday-at-the-seaside comedy); however, given the staged nature of action in studio films and the editing that yields shifts between individuals and scenes, such films of course depict something rather different from naturalistic human action. The events Avrahami and Kareev employed were also unnatural in the sense that the clipping and recombining procedure they used to generate stimuli would seem to have created physically impossible junctures between event units. A different issue is that the manipulations Avrahami and Kareev introduced may have enhanced grouping in the measures they report because these manipulations affected psychological meaningfulness rather than statistical learning, per se. Gaining evidence for statistical learning was not Avrahami and Kareev's express purpose, and thus they did not introduce control procedures specifically to counter the role of meaningfulness, as we did in the present research. Finally, for much the same reason, Avrahami and Kareev did not employ the statistical-learning techniques (e.g., Hunt & Aslin, 2001; Saffran et al., 1997) that have become standard in recent years. All in all, it is difficult to know both whether their findings (a) generalize to processing of everyday human action and/or (b) reflect the operation of statistical learning mechanisms in any way comparable to those that appear to operate in language and other aspects of visual processing.

2. Experiment 1

Experiment 1 provided a first test of whether sensitivity to statistical regularities within a novel string of small-scale acts (henceforth called "motion elements") enabled adults to "bind together" regularly co-occurring motion elements into higher-level, coarse-grained segments (henceforth called "actions") within the con-

tinuous flow of motion. To do this, we modified the recently innovated statistical-learning methodologies that were developed to test this issue in the language domain (e.g., Saffran et al., 1996, 1997). The basic approach adopted in such research has been to present observers with novel sequences of syllables (e.g., *go, la, bu, tu, pi, ro, bi, da, ku, pa, do, ti*) in which statistical regularities, supplying the only available clue to segmentation, were directly manipulated. For example, the sequence would be composed of four tri-syllabic combinations (e.g., *go-la-bu*, *tu-pi-ro*, *bi-da-ku*, and *pa-do-ti*) that were randomly intermixed across the exposure corpus, meaning *la* followed *go* and *bu* followed *la* with a transitional probability of 1.0, whereas *tu* followed *bu* with an average transitional probability of only .30.

Bringing this methodology to the action domain involved creating digitized video clips of twelve different motion elements. In constructing these motion elements, we opted to utilize everyday intentional actions, filmed in an uncluttered context with only a few everyday objects. To this end, each of the twelve different motion elements (e.g., *pour, poke, clink*) we employed involved action on a bottle plus, in some cases, another nearby object. These motion elements could be recombined in any sequence with the resulting stream of activity seeming fairly natural (no physical or bodily laws or constraints were violated). To achieve this, each clip began and ended with the actor's body in the same position. Via random selection, four three-motion-element combinations were selected (henceforth to be called "actions"), and these "actions" were then randomly intermixed to create a continuous, silent 20 min digitized video (henceforth to be called the "exposure corpus").

Motion within the exposure corpus was continuous. However, the individual motion elements (e.g., *pour, poke, clink*) within the exposure corpus were familiar to observers and readily extractable as segments (just as the syllables in the language research described earlier were presumed to be readily segmentable and familiar to both adult and infant listeners). At the same time, *combinations* of these motion elements were novel to observers and held no obvious causal or intentional significance. To illustrate, a sequence in which (a) the actor scrubbed the base of the bottle on a sponge, then (b) inserted the bottle into a nearby glass, and then (c) took a drink from the bottle, had no obvious meaning from a causal or intentional standpoint at the level of the motion-element triad. Again, then, the *sequence* of motion elements in the exposure corpus was both novel and not inherently meaningful (and recall that co-occurring motion elements were randomly selected). These features of the exposure corpus were a crucial aspect of our methodological approach: the novelty and relative meaninglessness of the "actions" (predictably co-occurring triads of motion elements) rendered adults unable to utilize any preformulated knowledge or inferences about intentions and goals to discover these "action" segments within the continuously streaming exposure corpus. Thus sequential probabilities across the motion elements within the stream supplied the only available clue for identifying higher-level "action" segments. Although we are inclined to think that, in everyday life, pre-existing conceptual knowledge of intentions, goals, and causes contributes to action segmentation, it was important to rule out influence of such causal/intentional knowledge on action segmentation within the context of these experiments in order to definitively demonstrate a possible role for statistical learn-

ing in adults' initial discovery of "action" segments within the stream of dynamic activity.

One additional methodological control was included to further ensure that inherent causal/intentional meaning or significance of the motion-element triads (the "actions") was not assisting adults in their discovery of actions within the exposure corpus. We created two different exposure corpora, with participants experiencing just one or the other of these corpora. In particular, the combination of motion elements selected to co-occur (i.e., the "actions") in one exposure corpus served as the foil combinations (e.g., the "non-actions," see detailed description below) in the other exposure corpus. If participants during a test phase could successfully discriminate "actions" from "non-actions," regardless of which exposure corpus they encountered, this would provide clearcut reassurance that the ability to identify "actions" was a result of sensitivity to statistics within the exposure corpus, and not a result of any inherent causal/intentional meaningfulness of the "actions."

The statistical learning methodology we employed enabled us to examine whether sensitivity to sequential probabilities alone enables adults to discover segments of which they had no prior knowledge within a novel sequence of motion elements. Findings from this research would not directly clarify that adults deploy a statistically-based segmentation strategy in their *on-line* processing of human action (as is also a limitation in research documenting statistical learning in the language domain). This is an interesting issue for future research. That said, this research had the potential to provide the first direct documentation to date that structural knowledge has the potential to support segmentation of dynamic human action.

2.1. Methods

2.1.1. Participants

Twenty-four undergraduates (12 female and 12 male, with equal numbers of females and males experiencing each of the stimulus sets) received course credit for participation in the research.

2.1.2. Materials

2.1.2.1. Exposure corpus. A given participant viewed an exposure corpus comprised of multiple repetitions of 12 video clips, with each clip depicting some kind of small motion (e.g., *stack, poke, drink*) involving a blue bottle, green glass, and/or yellow sponge. Each video clip began and ended with the actor and objects in virtually the same position (as close to identical as a human actor could achieve), enabling us to link any video clip with any other video clip. Video clips ranged in length from 2.2–4.3 s, $M = 3.3$ s at their actual filmed rate. Fig. 1 displays still images that depict just one frame from the video clip of each small motion-element (for illustrative purposes we selected the particular still frames that, in our judgment, best captured the unique identity of each motion element, but participants in the research of course viewed the full video clips from which these still frames were extracted).

Via random selection we created four three-motion "action" combinations (e.g., "actions" in one set (A) were *stack-poke-drink*, *blow-touch-rattle*, *pour-inspect-peek*,

Fig. 1. Illustrative still-images extracted from each video clip utilized in the exposure corpus across all sets in Experiments 1 and 2. The four rows of three images (viewing left to right) depict the four "actions" from set A.

insert-clink-scrub). As noted earlier, the use of random selection meant that the "actions" created via the three-clip combinations were no more inherently sensible nor familiar than any other possible combinations of video clips from the twelve-motion inventory we were working with. This was an important methodological feature of the research, ensuring that participants would have no prior causal/intentional knowledge or expectancies that could assist them in discovering action segments.

The four "actions" were then ordered randomly (with the stipulation that the same "action" never recurred in immediate succession) to create the 20 min exposure corpus. Each "action" occurred just 28 times (with 112 "actions" overall) in the exposure corpus. The only clues to "action" boundaries within each exposure corpus were the sequential probabilities between adjacent motion elements, which were higher *within* "actions" (1.0 in all cases; for example, *stack-poke*) than *between* "actions" (averaging 0.33; for example, *drink-pour*). Fig. 2 displays still frames selected to illustrate the flow of behavior within a small portion of the exposure corpus in one set (A).

Fig. 2. Illustrative still-images depicting the flow of events within a small portion of the exposure corpus from set A used in Experiments 1 and 2. Images should be "read" from left to right beginning at the top row.

For all adjacent motion elements throughout the entire exposure corpus, Macintosh iMovie software (Apple, Inc., Cupertino, CA) "overlap" transitions were used to smooth the shifts from one video clip to the next. As described earlier, the actor's position at the beginning and end of each video clip was as close to identical as could be achieved (still frames in Fig. 2 illustrate the degree of similarity). The transitions helped to smooth across the small degree of discrepancy that remained, yielding the general sense of a flow of behavior. The identical transition was used to link all video clips, regardless of whether the clips were part of the same "action" or not. Thus the quality of transitions themselves did not provide any information about segment boundaries.

If the exposure corpus were to begin at an "action" onset and/or end at an "action" offset, this in itself would provide a possible clue to segment boundaries. For this reason, in all cases the exposure corpus began and ended with a segment-internal small motion-element.

Two exposure corpora, sets A and B, were created using the method thus far described. As described briefly earlier, the only difference between the two corpora was the ordering of video clips within "actions"; in particular, sequences that served as "actions" in set A were "non-actions" in set B, and vice versa (see next section regarding test stimuli for further details). This set manipulation was another important design feature that helped to ensure that adults' ability to extract and recognize "action" segments arose from statistical regularities embedded in the exposure corpus and did not stem artifactually from fortuitous naturalness or meaningfulness of

some combinations of motion elements. The ordering of "actions" in sets A and B was yoked, such that each occurrence of a particular "action" in set A was linked to the occurrence of a yoked "action" in set B (e.g., set B "actions" were *touch-inspect-poke*, *rattle-peek-pour*, *blow-clink-stack*, *scrub-insert-drink* and *touch-inspect-poke* in set B occurred in the same position that *stack-poke-drink* occurred in set A).

2.1.2.2. Test stimuli. The purpose of the test stimuli was to probe adults' sensitivity to the statistical structure of the exposure corpus. Test stimuli consisted of 16 sequentially presented pairs, with each pair contrasting one "action" (e.g., as in the prior example: *stack-poke-drink*) with one "non-action." "Non-actions" were sequences containing motions adults had observed in the exposure corpus, but never in this particular combination (e.g., in set A, *touch-inspect-poke*, as depicted in Fig. 3). The "action" and "non-action" within a pair of test stimuli were separated by a 500 ms black screen.

Sequential probabilities between the motion elements in "non-actions" were all zero relative to the exposure corpus, because these pairings had never occurred. In contrast, sequential probabilities of video clips *within* "actions" were all 1.0. One test-videotape was generated in which a random ordering of the 16 pairs was selected, with half of these pairs depicting the "action" first and the "non-action" second. A second test videotape was created that maintained the same order of pairs, but reversed the ordering *within* pairs (e.g., if the "action" had preceded the "non-action" within a given pair in the first videotape, this "non-action" preceded the "action" for this pair in the second videotape). These counterbalancing precautions ensured that a possible response bias favoring the first or second clip within a pair could not be the source of an ability on adults' part to discriminate the "actions" from the "non-actions." The same two test videotapes employed for adults participating in set A were also employed for set B (given that "actions" in set A were

Fig. 3. Illustrative still-image examples of test stimuli in set A used in Experiments 1 and 2.

"non-actions" in set B, and vice-versa), with correct responding on a given test videotape corresponding to selection of opposite alternatives across the two sets.

2.1.3. Procedure

Participants in all experiments were asked to watch the exposure corpus and told they would subsequently be tested for memory of what they saw, but they were not given any indication regarding which aspects of the corpus the testing might concern. We opted to inform them about subsequent testing in order to ensure that they were motivated to watch the full exposure corpus (which was lengthy and repetitive). After viewing the exposure corpus they then were given several response-training trials, orienting them to use of the response sheet to be used in the test phase. These training trials involved paired action sequences depicting actions entirely unrelated to those depicted in the exposure corpus. In the test phase proper, adults were asked to select the member of each pair that they remembered having seen on their previous viewing. Half of the participants (12) viewed the set A exposure corpus, and half the set B exposure corpus. Within each set, half of the participants (6) provided responses with each of the two test videotapes.

2.2. Results and discussion

Collapsing across sets and test videotapes, adults discriminated "actions" from "non-actions" 80% ($SD = 23$) of the time, a rate significantly greater than predicted by chance (one-sample $t(23) = 6.4$, $p < .0001$) (see Fig. 4).

Eighteen of the 24 participants selected "actions" more frequently than "non-actions" (i.e., on nine or more of the 16 test trials), which was significantly greater than chance by a binomial test, $p < .01$. Performance was significantly better for

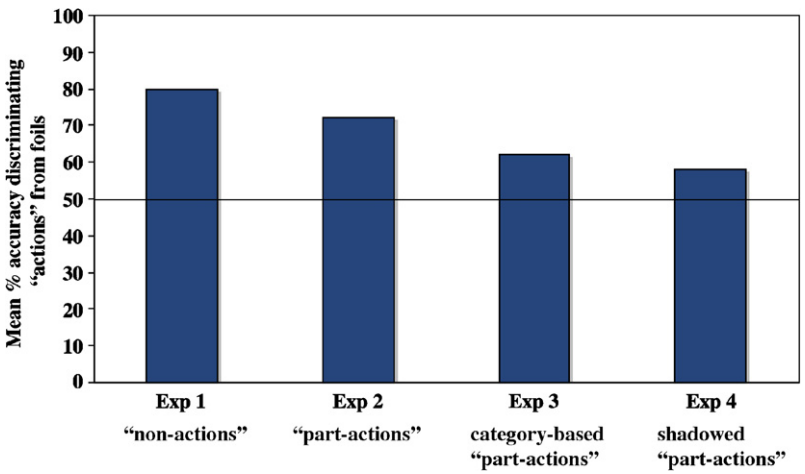

Fig. 4. Participants' mean percent accuracy in discriminating "actions" from foils (either "non-actions" or "part-actions", depending on the experiment) across all four experiments.

set A than set B ($t(22) = 2.3$, $p < .05$). However, when each of the sets was considered separately, in each set adults discriminated "actions" from "non-actions" at greater-than-chance rates, paired t's(11) = 6.9 and 3.1, one-tailed p's < .001 and .01 for sets A and B, respectively. Given the counterbalancing across sets, adults' selection of "actions" could not have arisen solely from greater salience of, or any general preference for, those sequences relative to the "non-action" sequences.

These findings confirmed that adults detect sequential probabilities among motion elements within a novel sequence of intentional action based on relatively brief exposure. However, a more stringent test of adults' skill at tracking segmentation-relevant sequential probabilities within the same stimulus stream could be undertaken: one in which adults are asked to discriminate "actions" (recurring sequences of motion elements) from combinations of small motion-elements that span action boundaries (called "part-actions"). This would be a more stringent test of adults' ability to detect statistical regularities within the exposure corpus because, unlike "non-actions," participants in fact encountered "part-action" combinations within the exposure corpus. If adults indeed can utilize sequential probabilities as a clue to action segmentation, "part-actions" should be perceived as violating the segmental structure they had discovered within the exposure corpus, and thus we predicted that they should select "actions" rather than "part-actions" on the post-hoc discrimination task. Experiment 2 tested this prediction.

3. Experiment 2

The second experiment investigated adults' ability to distinguish "actions" from recurring sequences of motion that *span* action boundaries but have lower sequential probabilities. To take a real-world example, consider a motion sequence in which someone grasps a cream pitcher, pours cream into coffee and then fluidly moves to pick up a telephone receiver to make a phone call. In this example, discriminating "actions" from "part-actions" would involve distinguishing the sequence *grasp cream-pitcher/pour cream* (constituting the "action" segment *cream coffee*) from the sequence *pour cream/grasp telephone* (a "part-action" spanning portions of the *cream coffee* and *phone call* segments). As in Experiment 1, participants viewed one of two exposure corpora prior to being tested on their ability to discriminate "actions" from "part-actions," with the "actions" in one exposure corpus serving as "part-actions" in the other, and vice versa.

3.1. Methods

3.1.1. Participants

Twenty-four undergraduates (12 female and 12 male, with equal numbers of each gender experiencing each stimulus set) received course credit for participation in the research.

3.1.2. Materials

3.1.2.1. Exposure corpus. Adults viewed an exposure corpus in one of two counterbalanced sets (A and C), with "actions" in the exposure corpus of set A serving as "part-actions" in the exposure corpus of set C, and vice versa. "Actions" in set A were *stack-poke-drink, blow-touch-rattle, insert-clink-scrub,* and *pour-inspect-peek,* and in set C were *drink-blow-touch*, *clink-scrub-pour*, *rattle-stack-poke*, and *inspect-peek-insert*.

3.1.2.2. Test stimuli. Test videotapes were constructed as in Experiment 1, except that pairs of test stimuli contrasted "actions" with "part-actions" (sequences that participants had actually seen in the exposure corpus that included three video clips spanning "action" boundaries; in set A, for example, *drink-blow-touch*, as depicted in Fig. 3). In particular, two videotapes of 16 pairs of test stimuli were constructed (used in both sets), this time exhaustively pairing the four "actions" with four "part-actions." Across the two videotapes, the order of "actions" versus "part-actions" within test pairs (i.e., first vs. second) was counterbalanced.

Within "part-actions," one pair of adjacent clips had average sequential probabilities of 0.33, and the other two adjacent clips had sequential probabilities of 1.0, yielding an overall average co-occurrence frequency of 0.67 for "part-actions," versus an overall average co-occurrence frequency of 1.0 for "actions."

3.1.3. Procedure

The procedure was comparable to that of Experiment 1 except that test item pairs involved discrimination between "actions" and "part-actions" (rather than "actions" and "non-actions" as in the first experiment).

3.2. Results and discussion

Discriminating "actions" from "part-actions" ought to be more challenging than the discrimination task in Experiment 1 regarding "actions" versus "non-actions" because overall sequential probabilities for video clips in "part-actions" were substantially greater than zero. Despite the potential difficulty of the discrimination task, adults systematically selected "actions" more frequently than "part-actions" ($M = 72\%$, $SD = 24$; one-sample $t(23) = 4.5$, $p < .001$). This level of performance did not differ significantly from discrimination of "actions" versus "non-actions" in the first experiment (see Fig. 4). As well, as in Experiment 1, 18 of the 24 participants in Experiment 2 selected "actions" more frequently than "part-actions" (i.e., on nine or more of the 16 test trials), which was significantly greater than chance in a binomial test, $p < .01$. No significant difference between sets A and C emerged in discriminating "actions" from "part-actions." In sum, even though low-frequency adjacencies were presented as part of the same "unit" in the test-stimuli (given the bracketing of "part-actions" from "actions" by 500 ms black screens within a given test pair), participants were disinclined to select these "part-action" units in the test phase. Instead, test-stimuli including high-frequency adjacencies that had been bounded by low-frequency adjacencies in the exposure corpus ("actions") were selected at relatively high rates.

We predicted that adults would be able to discriminate "part-actions" from "actions" because "part-actions" should be perceived as violating the segmental structure that adults had discovered within the exposure corpus via tracking of sequential dependencies. That said, adults' skill at discriminating "part-actions" from "actions" can be accounted for by their having detected joint probabilities of just two adjacent small motion-elements. In other words, adults need not necessarily have extracted the full three-element *triads* to have succeeded at the task. At the very least, then, the Experiment 2 findings clarify that adults utilized sequential probabilities to extract segments composed of two adjacent small motion-elements. Additional probes, such as those pioneered by Aslin and colleagues (e.g., Aslin, Saffran, & Newport, 1998; Fiser & Aslin, 2002a, 2002b) in domains including language, tone sequences, shape sequences, spatial arrays, and visuomotor tasks, will be needed to clarify whether extraction of full triads or of higher-order statistics such as conditional probability play a role in the statistical learning adults display in action processing. In any case, however, the findings of Experiment 2 clarify that adults succeeded in detecting and remembering particular groupings of small motion-elements. Such sensitivity (a) gives them a basis on which to segment novel action scenarios for which no other clues to segmentation are on offer, and (b) might assist their segmentation in everyday action processing.

4. Experiment 3

While interesting, findings from Experiments 1 and 2 do not warrant ready generalization of adults' statistically-based segmentation skills to the processing of everyday action, in part because the action sequences used in those experiments involved repeated viewing of identical instances. In the real-world context, of course, instances of any given action sequence are rarely, if ever, identical. Rather, everyday intentional action exhibits substantial variability: Across occasions, for example, a grasping motion is performed at varying rates, with different hands, on different objects in different locations with respect to the body, and with different hand/finger configurations. It will ultimately be important to determine how effectively statistical learning guides segmentation in the face of such variability.

One step to take in bringing the statistical learning paradigm closer to accounting for real-world processing is to examine whether adults can detect statistical structure across *categories* of similar, but not identical, instances of small motion-elements. Statistical learning for simple visual shapes (e.g., Turk-Browne et al., 2005) and at least some aspects of language (e.g., Gerken, Wilson, & Lewis, 2005; Gomez & Gerken, 1999; Thompson & Newport, 2007) encompasses this level of skill, making it seem plausible for the action domain as well. Experiment 3 thus tested whether adults can discover such category-based "actions" and subsequently discriminate them from category-based "part-actions" (combinations of small motion-elements that span the boundary between two "actions" in the exposure corpus). Such a finding would further bolster the plausibility of statistical learning as a mechanism subserving segmentation of everyday action.

4.1. Methods

4.1.1. Participants

Seventeen undergraduates (nine male and eight female; set A: four female, five male; set C: three female, five male) received course credit for participation in the research. Data from one participant was eliminated due to being a clear outlier (performance was different by more than 2.5 standard deviations from the mean of the other participants). This participant's data were replaced by that of the seventeenth participant.

4.1.2. Materials

4.1.2.1. Exposure corpus. As in Experiment 2, adults viewed an exposure corpus in one of two counterbalanced sets (A and C), with "actions" in the exposure corpus of set A serving as "part-actions" in the exposure corpus of set C, and vice versa. A new randomized selection of small motion-elements to create "actions" and "part-actions" was undertaken in order to increase generalizability while maintaining all other facets of the structure of the exposure corpus employed in Experiment 2. "Actions" in set A of Experiment 3 were *blow-insert-drink, touch-poke-pour, peek-clink-rattle,*and *scrub-inspect-stack,* and in set C were *rattle-blow-insert*, *drink-peek-clink*, *inspect-stack-touch*, and *poke-pour-scrub*. The primary methodological change in Experiment 3 was that 28 different versions of each small motion-element were filmed. Each time a small motion-element appeared in the exposure corpus, it was a new version relative to all previously viewed versions. Thus, the sequential probabilities of small motion-elements represented by "actions" and "part-actions" were the same as in Experiment 2 (an overall average of 0.67 for adjacent small motion-elements within "part-actions", and 1.0 for "actions"), but these probabilities had to be tracked across *categories* of different versions of each small motion-element.

The different versions of each small motion-element differed along a variety of dimensions such as rate of motion, position of hand on the object(s), which hand(s) was/were involved, overall body configuration, path of motion, and manner of motion. Each of the 28 versions of all twelve small motion-elements were filmed on the same occasion, making the lighting conditions and positions of objects constant across all small motion-elements utilized in the experiment. Our twofold goal in constructing the 28 versions of each small motion-element was to (a) introduce enough variability among the different versions of a given small motion-element that observers would be able to readily discriminate different versions from one another while at the same time (b) maintaining enough similarity across versions that observers would be able to group the different versions of a given small motion-element into a single small-motion category. To illustrate the flavor of such variability, Fig. 5 depicts several of the different versions utilized for four of the twelve small motion-elements.

If adults were to succeed in tracking sequential probabilities to discover segmental structure within the exposure corpus in Experiment 3, this would indicate that our second goal was met. To document that the first goal was met, however,

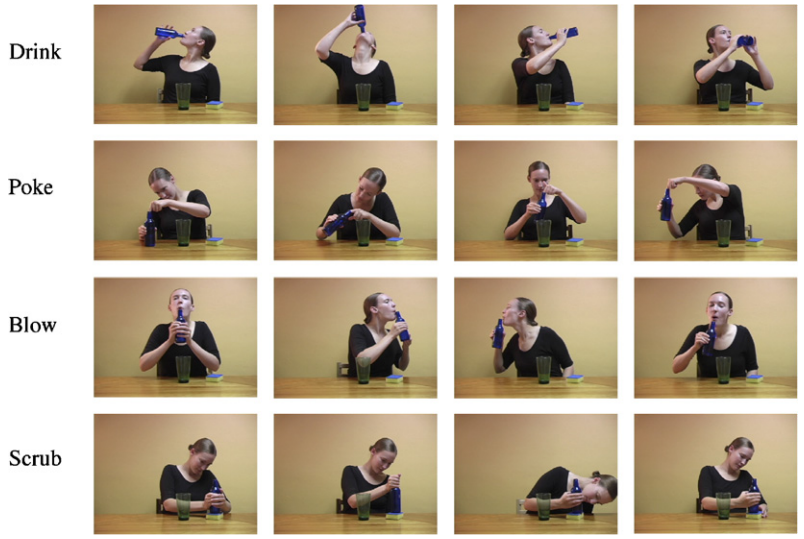

Fig. 5. Illustrative still-image examples of different versions of four of the small motion-element categories utilized as stimuli in Experiment 3.

we collected discrimination data. Adults were shown pairs of video clips in sequence and asked to judge whether the two clips in a pair were the same or different. Each pair depicted either (a) two different versions of a given small motion-element or (b) the identical small motion-element video clip. Collecting discrimination data for all possible same and different pairings of all 28 versions of all 12 small motion-elements would be a prohibitively lengthy task for participants. We thus randomly selected 10 video clips from each of the 28 versions of each of the 12 small motion-elements to include in the discrimination task. These 10 video clips enabled us to include five "Different" trials for a given small motion-element, and five "Same" trials were also constructed for each of the twelve small motion-elements. In all, adults made same/different judgments on 120 trials across all of the 12 small motion-elements (10 judgments for each of the 12 small motion-elements). As predicted, adults readily discriminated the different versions: they displayed an overall mean same/different accuracy rate of 94.3% ($SD = 5.5$), with a mean accuracy of 91.8% ($SD = 11.6$) for same trials and 96.8% ($SD = 3.3$) for different trials. Mean accuracy rates across the 12 different small motion-elements were uniformly high (range: 87% ($SD = 4.8$) to 97% ($SD = 4.8$).

4.1.2.2. Test stimuli. Test videotapes were constructed as in Experiment 2 (16 pairs of test stimuli contrasted all possible combinations of the four "actions" with four "part-actions") except that each of the four "action" test stimuli were comprised of different versions of the relevant small motion-elements, as were each of the four "part-action" test stimuli. The small motion-elements utilized in each of

the test stimuli were randomly selected from among those participants had seen during their viewing of the exposure corpus. Thus the test stimuli probed participants' ability to track sequential dependencies across categories of small motion-elements, but, strictly speaking, did not test for their ability to generalize this detection to an entirely new set of exemplars of those categories of small motion-elements. As in Experiment 2, the order of "actions" versus "part-actions" within test pairs (i.e., first vs. second) was counterbalanced across the two test videotapes.

4.1.3. Procedure

The procedure was comparable to that of Experiment 2, with two exceptions. Because of the considerably greater difficulty of the statistical learning task involved in tracking sequential probabilities across categories of small motion-elements, participants were given a larger sample of the statistics before being tested. This was accomplished by having them view the exposure corpus twice prior to test. To avoid participants' attention lagging during viewing of the doubled exposure corpus, the presentation rate of the action stimuli was speeded, resulting in doubled viewing of the exposure corpus lasting just over 20 min.[3]

4.2. Results and discussion

Adults were able to discover "action" segments within the Experiment 3 exposure corpus despite being faced with discriminably different versions of each of the small motion-elements comprising those "actions": In the post-test they selected "actions" more frequently than "part-actions" 62% of the time ($SD = 17$), a level significantly greater than predicted by chance, one-sample $t(15) = 2.8$, $p < .05$[4] (see Fig. 4). Twelve of the 16 participants selected "actions" more frequently than "part-actions" (on 9 or more of the 16 test trials), which was significantly different from chance by the binomial test, $p < .05$. No significant difference between sets A and C emerged in discriminating "actions" from "part-actions" in this category-based statistical learning task.

[3] Action speed was doubled via Macintosh iMovie software. Speeding of action resulted in action that appeared rapid but not particularly unnatural. This was in part because action at the normal rate of filming occurred at a leisurely pace. Prior to employing a speeded exposure corpus in Experiments 3 and 4, participants' ability to cope with speeded action was investigated employing the same stimuli that were used in the first two experiments. Participants displayed the same level of skill at discriminating "actions" from "non-actions" and "part-actions" with the speeded exposure corpus as with the exposure corpus presented at the normal filming rate, indicating that the speeded action presented little difficulty to processing.

[4] As described earlier, data from one participant in Experiment 3 were eliminated due to this participant's clearly outlying responses (an accuracy rate more than 2.5 standard deviations below the mean). When this participant's responses were included in parametric analyses, overall accuracy rates ($M = 58\%$, $SD = 22$) were no longer significantly greater than chance $t(16) = 1.54$, $p = .14$). Importantly, however, when including this participant's data, accuracy in selecting "actions" over "part-actions" remained significantly greater than predicted by chance in a non-parametric binomial test ($p < .05$), which is not subject to the influence of extreme scores.

The Experiment 3 findings clarify that adults are able to track statistical regularities across small motion-elements exhibiting a substantial degree of surface variability. This finding thus increases the plausibility that the segmentation-relevant statistical learning skill tapped in these studies could cope with real-world complexity.

5. Experiment 4

One additional important question concerning the findings of these experiments is whether they genuinely index sensitivity to statistical structure within *action*. It is conceivable that adults recoded the video clips into verbal labels – much as we have done in labeling the clips in our examples – and tracked statistics across these linguistic elements rather than across the video clips themselves. Of course, this would be a new and interesting finding in its own right, but it would not be evidence for an ability to capitalize on segmentation-relevant statistical structure in dynamic action, per se. To address this issue, we carried out an additional experiment parallel to Experiment 2 (''actions'' versus ''part-actions'') with one addition: Adults were asked to carry out a demanding linguistic task – verbally ''shadowing'' a story they were hearing over headphones (repeating words aloud as quickly as possible after hearing them; Cherry, 1953) – during the entire time they viewed the exposure corpus of dynamic action. Prior research has documented that shadowing strongly interferes with the recoding of visual information into verbal form (e.g., Besner, Davies, & Daniels, 1981; Estes, 1973; Levy, 1971; Murray, 1967; Posner, Early, Reiman, Pardo, & Dhawan, 1988; Wood & Cowan, 1995). The shadowing technique continues to be used in current research to isolate processes such as the transfer of information to visual (as opposed to verbal) working memory (e.g., Schmidt, Vogel, Woodman, & Luck, 2002). The shadowing task selected for the present research was a very demanding one – shadowing a rapidly unfolding narrative delivered by a narrator speaking with a pronounced British accent (which presented a unique challenge to our North American participants). Thus, adults who were shadowing while watching the exposure corpus should lack the attentional resources needed to phonologically recode the motion elements in that corpus into verbal labels.

5.1. Methods

5.1.1. Participants

Twenty-four undergraduates (17 female, 7 male; set A: 8 female and 4 male; set B: 9 female and 3 male) received course credit for participation in the research.

5.1.2. Materials

Exposure corpora and test stimuli employed were comparable to those in Experiment 2 involving discrimination between ''actions'' and ''part-actions,'' except that the small motion-element combinations comprising ''actions'' and ''part-actions'' were those utilized in Experiment 3 (that is, ''actions'' in set A were *blow-insert-drink,*

touch-poke-pour, peek-clink-rattle, scrub-inspect-stack, and in set C were *rattle-blow-insert*, *drink-peek-clink*, *inspect-stack-touch*, and *poke-pour-scrub*).[5] The video clips used were the same as those in Experiment 2. The shadowed story employed in Experiment 4 was "Charlie and the Chocolate Factory" narrated by the author, Roald Dahl, speaking in British-accented English (Dahl, 1975). The mean narration rate for the story was 163.5 words per minute.

5.1.3. Procedure

The same procedure was employed as in Experiment 3, except that adults shadowed a story while viewing the exposure corpus. In particular, as in Experiment 3, participants viewed the exposure corpus twice to compensate for the increased challenge introduced by shadowing. Regarding the shadowing, participants were asked to repeat words as quickly as they could after hearing them. Participants' shadowing was audiotaped and subsequently double-checked to ensure that it occurred throughout viewing of the exposure corpus.

5.2. Results and discussion

As in our previous experiments, those who carried out the shadowing task while viewing the exposure corpus selected the "actions" ($M = 58\%$, $SD = 15$) in the test phase significantly more frequently than predicted by chance (one sample $t(23) = 2.3$, $p < .05$) (see Fig. 4), with no significant difference in performance between sets A and C. Seventeen of the 24 participants selected "actions" more often than "part-actions" in the test phase (i.e., on nine or more of the 16 test trials), which was significantly different from chance by a binomial test, $p < .05$. Participants' absolute level of "action" selection was significantly lower than that of adults in Experiment 2 who did not shadow ($M = 72\%$, $SD = 24$), independent samples $t(46) = 2.7$, $p < .05$, but this is unsurprising given that shadowers were faced with carrying out two challenging processing tasks simultaneously while non-shadowers contended with only one of these tasks. Others have documented analogous decrements in detection of statistical regularities in a dual-task context in the language domain (e.g., Toro, Sinnett, & Soto-Faraco, 2005) and the visual domain (e.g., Turk-Browne et al., 2005). The crucial finding for present purposes is that shadowers were sensitive to sequential probabilities within the motion stream that provide clues to segmentation even when linguistic resources for recoding of small motion-elements into verbal labels were unavailable to assist in extracting action segments. Their success in dis-

[5] One error was belatedly detected in the exposure corpus observed by participants in Condition C of Experiment 4. In particular, on two of the 28 occasions when participants were supposed to view the "action" *poke-pour-scrub* they were mistakenly shown *scrub-pour-poke* instead. Both of these errors occurred relatively early in the exposure corpus (the 7th and 14th "actions" viewed in the series). The exposure corpus of condition A did not include this error. Interestingly, this error should have made the "actions" even more difficult to detect than had no such error been present, yet participants were able to discriminate the statistically-based "actions" from "part-actions" despite this inadvertently-introduced random variability.

criminating ''actions'' that were solely statistically defined was striking given that the statistics involved had to be tracked while they were simultaneously dealing with the challenging shadowing task.

In sum, Experiment 4 clarified that recoding of small motion-elements into verbal labels was not the sole source of adults' success at tracking statistics within the stream of dynamic action. Rather, adults directly detected statistical structure within dynamic intentional action and used it as a source of information for discriminating higher-level segments (''actions'') from motion combinations that violated the segmental structure of the behavior stream (''part-actions'').

6. General discussion

The experiments reported here confirm that adults can discover sequential probabilities within dynamic intentional activity that support extraction of higher-level action segments. Adults displayed such skill even in the face of substantial surface variability in the low-level segments over which statistics were tracked, and they accomplished this even when use of a linguistic recoding strategy was drastically undercut. These studies offer among the first pieces of evidence illuminating a potential mechanism underlying segmentation of dynamic human action.

To ensure that sequential probabilities were the only possible basis for segmentation, we took steps (e.g., use of digitized video and a standard transition from one video clip to the next, random selection of motion element combinations in the construction of the exposure corpus) to systematically eliminate both (a) top–down sources of information about intentions, goals, and causes that might guide discovery of higher-level segments, as well as (b) other possible structural clues to coarse-grained segmentation of the motion stream except sequential probabilities. Furthermore, we included an important control that confirms that top–down knowledge of intentions, goals, and causes was not the source of adults' ability to discover the ''actions'' within the exposure corpus. That is, the motion-elements comprising ''actions'' that one group of participants had the opportunity to extract via sequential probabilities served as the foil ''non-actions'' (Experiment 1) or ''part-actions'' (Experiments 2-4) that other participants encountered. Participants' ability to extract the ''actions'' regardless of which set of motion elements were involved clarifies that sequential probabilities drove accurate selection of ''action'' segments over ''non-actions'' or ''part-actions''; inherent causal/intentional meaningfulness of the motion element combinations comprising those ''actions'' could not have been the source of adults' accuracy. Thus the present findings specifically demonstrate adults' ability to exploit sequential probabilities that enable discovery of higher-level segments not otherwise detectable within a stream of behavior.

It is noteworthy that adults in the present research were given access to only a small sample of the segmentation-relevant statistics (in Experiments 1 and 2 only 28 exposures to each ''action'' relative to 80 or more exposures to the relevant segments in analogous language segmentation research). Nevertheless, they accurately

discriminated ''actions'' from ''non-actions'' and ''part-actions,'' revealing clearcut sensitivity to the sequential probabilities inherent in the exposure corpus.

Our findings should in no way be taken to indicate that action segmentation in the every day setting reduces to statistically-driven detection of segment boundaries. Rather, we believe that action segmentation in the real world is likely the joint product of numerous mechanisms – both knowledge of intentions, goals, causes, and action propensities as well as a suite of other mechanisms involving structural knowledge and sensitivity to bottom–up clues. However, by eliminating the possible operation of such other mechanisms, the present research provides the first direct evidence that statistical learning is one mechanism adults can exploit to facilitate action segmentation. That said, an important direction for future research will be to investigate how mechanisms such as statistical learning may be deployed in richer, more meaningful real-world contexts to subserve action processing.

The statistical learning paradigm we employed in these studies provides evidence that adults can track sequential probabilities relevant to initial discovery of higher-level segments within a novel sequence of activity. These findings do not, however, directly clarify whether adults utilize sequential probabilities to drive their on-line segmentation of dynamic human action. This remains a question for future investigation. Techniques pioneered by others for investigating this question in other domains (e.g., Hunt & Aslin, 2001; Olsen & Chun, 2001;Turk-Browne et al., 2005) can be brought to bear in future research investigating statistical learning in the action domain. That said, Swallow and Zacks' (submitted for publication) recent work increases the plausibility that statistical learning may benefit on-line segmentation in the action domain. They showed adults arbitrary sequences of animated still-frames depicting novel hand gestures, and found that sensitivity to statistical regularities within the sequences guided adults' on-line attentional deployment during processing. Although Swallow and Zacks' studies did not involve actual dynamic action, their findings clearly document that statistical learning with action-approximating static stimuli have implications for on-line processing.

Our findings and those of Swallow and Zacks nicely complement one another in yet another respect. In contrast to Swallow and Zacks, the intentional action sequences we utilized in the present research involved motion elements that are *non-arbitrary*, in the sense that observers would have a sense of the purpose or goal of each individual motion element (it was the *combinations* of these elements that were novel and not inherently meaningful). Thus, together the present findings and those of Swallow and Zacks clarify that statistical learning supports observers' extraction of more coarse-grained segments from within a behavior stream, irrespective of the arbitrariness of the motion elements comprising that motion stream.

6.1. Coping with real-world complexity

One of the strengths of the present research is the use of video of an actual human carrying out a physically possible yet novel sequence of everyday intentional action. These findings thus convincingly extend statistical learning skills to the realm of dynamic action stimuli. At the same time, there are significant reasons for caution

at this early phase in generalizing the present findings to segmentation of everyday action in the real-world. For one, it is important to recognize that, in everyday action, statistical regularities are likely probabilistically supplemented by other structural clues to action segments, such as acceleration and deceleration regularities that coincide with segments (e.g., Loucks & Baldwin, 2006; Zacks, 2004), increases in movement change (e.g., Hard et al., under review; Newtson et al., 1977), and occasional pauses. In particular, research by Hard and colleagues (Hard et al., in press) indicates that movement change is greater at points within the motion stream that observers identify as boundaries between distinct acts. Moreover, in the motion scenarios they have investigated, movement change appears to be greatest at boundaries between coarse-grain action segments relative to fine-grained segment boundaries. These findings hint that observers may track movement change to help in detecting action segments, and possibly even to guide hierarchical organization nesting fine-grained action segments within coarse-grained segments. However, the extent to which adults actually rely on such additional clues is not yet known. Investigation of this question and other such questions (e.g., how the variety of structural clues are integrated with one another, and with expectancies based on causal/intentional knowledge) are important avenues for future research.

Statistical regularities likely abound in dynamic intentional action, and this research confirms adults' sensitivity to such structure. Yet very little is known about details of the actual regularities to be found within everyday intentional action. Ultimately, it will be important to characterize the statistical structure in the natural world in order to fully understand how effectively statistical learning-skills can support segmentation in the context of real-world complexity. One thing is clear, however: real-world statistical regularities are more complex than those exhibited in the corpus of action adults observed in the present research. It is thus not entirely certain whether the statistical-tracking skills adults demonstrated in the present studies are powerful enough to support action segmentation in the context of real-world statistical complexity, or what size of learning corpus would be needed for segments to be extracted given such complexity. Three points alleviate this concern, at least to some degree. First, adults detected entirely novel and inherently meaningless motion combinations after observing only a small sample of these combinations, and they were able to do so even when engaged in linguistic shadowing, a resource-intensive cognitive/linguistic task. This points to statistical-tracking skills that are rapid and powerful. Second, Experiment 3 confronted adults with a substantially increased level of complexity, in that adults could discover higher-level "action segments" only if they could track sequential probabilities across categories of small motion-elements, the exemplars of which varied considerably on a range of surface characteristics. Adults readily accomplished segmentation in the context of such complexity despite the relatively small statistical sample.

Finally, another possible reason for caution in generalizing our findings to everyday action processing relates to the specifics of the action stimuli we presented to participants. For reasons already described, we constructed a stream of behavior in which small motion-elements all began and ended with the actor in the same, neutral position. Hence the return to neutral position potentially demarcated clearcut

boundaries between each of the small motion-elements. The stream of dynamic activity was thus readily segmentable for participants at the fine-grained level of the small motion-element, thereby facilitating their ability to detect sequential dependencies among these small motion-elements. While this design feature in no way invalidates the findings, one might question the extent to which everyday intentional action is similarly segmentable at the fine-grained level. If it is not, the statistical tracking mechanism we have showcased would have little opportunity to operate. Almost certainly, small motion-elements in the everyday setting do not come as neatly packaged by a single, salient, return-to-neutral boundary cue as the motion stream that participants experienced in this research. On the other hand, recent research suggests that fine-grained segments like the small motion-elements we presented to participants may indeed be highly available segments within intentional action, thus setting the stage for learning of sequential dependencies among such elements. For example, Zacks (2004) found that adults' segmentation judgments at the fine-grained level correlate especially highly with a cluster of movement features such as acceleration magnitude and speed. Moreover, when adults view intentional action in point-light format – which retains structure-in-motion but minimizes contextual information – their memory displays the influence of segmentation at the fine-grained level but not at the coarse-grained level (Baldwin et al., in preparation). Also, infants readily segment intentional action at the fine-grained level (Baldwin et al., 2001; Saylor et al., 2007) and can do so even when action is displayed in point-light format (Baldwin et al., in preparation). All in all, then, at this early phase there is at least some basis on which to speculate that the statistical tracking skills highlighted in the present research could well support the discovery of higher-level action segments in everyday action processing.

6.2. Broader implications

Quite possibly, sensitivity to statistical regularities within the human motion stream may also support aspects of action processing other than segmentation. For example, some have suggested that different types of motion/action, such as animate versus inanimate motion, intentional versus unintentional action, and diverse types of intentional acts (e.g., helping vs. hindering) (e.g., Blythe, Todd, & Miller, 1999; Mandler, 1988; Premack & Premack, 1995) are distinguishable via systematic structural differences. If this is correct, sensitivity to structural regularities could aid adults' rapid, automatic recognition of highly-relevant event distinctions within the motion stream as well as facilitating action segmentation.

The present findings highlight a striking parallel between intentional action processing and processing in other cognitive/perceptual domains, such as language. Using a methodology developed by Saffran and colleagues that demonstrated reliance on statistical structure for language segmentation, we obtained comparable findings that adults can segment novel action sequences via sequential probabilities. As briefly alluded to earlier, additional research documents similar statistical tracking in yet other domains, such as in processing of non-linguistic tone sequences and complex spatial arrays. Thus it seems likely that cognitive/perceptual systems for processing

language, action, and these other kinds of stimuli all recruit one and the same statistical tracking mechanism to facilitate segmentation. Many specific details concerning this domain-general learning mechanism remain as yet unresolved (e.g., Perruchet & Pacton, 2006). In any case, however, the current findings further underscore the importance of pursuing investigation of domain-general experience-dependent mechanisms that operate across distinct knowledge systems. At the same time, constraints on statistical computations may well emerge that are specific to the action domain or to certain aspects of action processing, as seems to be the case for statistical learning with linguistic stimuli (e.g., Bonatti, Peña, Nespor, & Mehler, 2005; Saffran & Thiessen, 2003) as well as simple visual shapes (e.g., Baker, Olson, & Behrmann, 2004).

6.3. Conclusion

Skill at identifying distinct acts within others' continuously flowing behavior is fundamental to everyday social and cognitive functioning. Yet action is complex – dynamic, evanescent, and largely continuous. How we accomplish segmentation of action in the face of such complexity is a basic question for cognitive science research. The experiments reported here demonstrate that adults can register statistical regularities that provide clues to action segmentation. In particular, adults tracked sequential probabilities among fine-grained actions (small motion-elements) to identify more coarse-grained segments of action within a novel sequence of intentional activity. This finding provides some of the first evidence to date about mechanisms that enable discovery of segments within dynamic human action, and, for the first time, documents a potential role for structural knowledge in action processing.

Acknowledgements

Our thanks to those who participated in this research, as well as to Amanda Altig, Stephen Boyd, Katherine Carlson, Alicia Craven, Kurstin Hollenbeck, Melissa Mason, Karen Myhr, and Catherine Tenedios for their assistance with data collection. This research was supported by the National Science Foundation under Grant No. BCS-0214484 to the first author and by funds from the University of Oregon.

References

Asch, S. (1952). *Social psychology*. Englewood Cliffs, NJ: Prentice Hall.

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Avrahami, J., & Kareev, Y. (1994). The emergence of events. *Cognition, 53*, 239–261.

Baird, J., & Baldwin, D. (2001). Making sense of human behavior. In B. Malle, L. Moses, & D. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 193–206). Cambridge, MA: MIT Press.

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15*, 460–466.

Baldwin, D., & Baird, J. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences, 5*, 171–178.

Baldwin, D. A., Baird, J. A., Malle, B., Neuhaus, E., Craven, A., Guha, G., & Sobel, D. (in preparation). Segmenting dynamic human action via sensitivity to structure in motion. Unpublished manuscript, University of Oregon.

Baldwin, D., Baird, J., Saylor, M., & Clark, A. (2001). Infants parse dynamic human action. *Child Development, 72*, 708–717.

Besner, D., Davies, J., & Daniels, S. (1981). Reading for meaning: The effects of concurrent articulation. *Quarterly Journal of Experimental Psychology, 33A*, 415–437.

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews: Neuroscience, 2*, 1–8.

Blythe, P., Todd, P., & Miller, G. (1999). How motion reveals intention: Categorizing social interactions. In G. Gigerenzer & P. Todd (Eds.), *Simple heuristics that make us smart* (pp. 257–285). Oxford, UK: Oxford Univ. Press.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science, 16*, 451–459.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical society of America, 25*, 975–979.

Cohen, A., Ivry, R. I., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 17–30.

Dahl, R. (1975). *The Roald Dahl audio collection*. Harper Children's Audio.

Estes, W. K. (1973). Phonemic coding and rehearsal in short-term memory for letter strings. *Journal of Verbal Learning and Verbal Behavior, 12*, 360–372.

Fiser, J., & Aslin, R. (2002a). Statistical learning of higher-order temporal structures from visual shape sequences. *Journal of Experimental Psycyhology: Learning, Memory and Cognition, 28*, 458–467.

Fiser, J., & Aslin, R. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99*, 15822–15826.

Frith, C., & Frith, U. (1999). Interacting minds: A biological basis. *Science, 286*, 1692–1695.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249–268.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135.

Hard, B. A., Lozano, S., & Tversky, B. (under review). Hierarchical encoding: A mechanism for observational learning. Journal of Experimental Psychology: General.

Hard, B. A., Tversky, B. & Lang, D. (in press). Making sense of abstract events: Building event schemas. Memory & Cognition.

Heider, F. (1958). *The psychology of interpersonal relations*. NY: Wiley.

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General, 130*.

Levy, B. A. (1971). Role of articulation in auditory and visual short-term memory. *Journal of Verbal Learning and Verbal Behavior, 10*, 123–132.

Loucks, J., & Baldwin, D. (2006). When is a grasp a grasp? Characterizing some basic components of human action processing. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets words: How children learn verbs* (pp. 228–261). New York: Oxford University Press.

Mandler, J. (1988). How to build a baby: On the development of an accessible representational system. *Cognitive Development, 3*, 113–136.

Murray, D. J. (1967). The role of speech responses in short-term memory. *Canadian Journal of Psychology, 21*, 263–276.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28*, 28–38.

Newtson, D., & Enquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology, 12*, 436–450.

Newtson, D., Enquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*, 847–862.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*, 1–32.

Olsen, I. R., & Chun, M. M. (2001). Temporal contextual cuing of visual attention. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 1299–1313.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences, 10*, 233–238.

Posner, M., Early, T., Reiman, E., Pardo, P., & Dhawan, M. (1988). Asymmetries in hemispheric control of attention in schizophrenia. *Archives of General Psychiatry, 45*, 814–821.

Premack, D., & Premack, A. (1995). Intention as psychological cause. In D. Sperber & D. Premack (Eds.), *Causal Cognition* (pp. 185–199). Clarendon Press.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Saffran, J., Newport, E., Aslin, R., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–105.

Saffran, J., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*, 484–494.

Saylor, M. M., Baldwin, D., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development, 8*, 113–128.

Schmidt, B. K., Vogel, E. K., Woodman, G. F., & Luck, S. J. (2002). Voluntary and automatic attentional control of visual working memory. *Perception and Psychophysics, 64*, 754–763.

Shiffrar, M., & Freyd, J. J. (1990). Apparent motion of the human body. *Psychological Science, 1*(4), 257–264.

Swallow, K., & Zacks, J. (submitted for publication). Sequences learned without awareness can orient attention during the perception of human activity.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3*, 1–42.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*, B25–B34.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*, 552–564.

Tversky, B., Zacks, J., & Martin Hard, B. A. (in press). T. Shipley & J. Zacks (Eds.) Understanding events: How humans see, represent, and act on events.

Wood, N. L., & Cowan, N. (1995). The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of Cherry (1953). *Journal of Experimental Psychology: General, 124*, 243–262.

Zacks, J. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*, 979–1008.

Zacks, J., Braver, T., Sheridan, M., Donaldson, D., Snyder, A., Ollinger, J., Buckner, R., & Raichle, M. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4*, 651–655.

Zacks, J., Swallow, K. M., Vettel, J. M., & McAvoy, M. P. (2006). Visual motion and the neural correlates of event perception. *Brain Research, 1076*, 150–162.

Zacks, J., & Tversky, B. (2001). Event structure in perception and conception. *Psychogical Bulletin, 127*, 3–21.