UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Event Segmentation in Language and Cognition

Permalink
https://escholarship.org/uc/item/6nm5b85t

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors
Tınaz, Bilge
Ünal, Ercenur

Publication Date
2024

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Event Segmentation in Language and Cognition

Bilge Tınaz (bilge.tinaz@ozu.edu.tr)
Özyeğin University, İstanbul, Turkey
Nişantepe Mahallesi, Orman Sokak, 34674, Çekmeköy, İstanbul, Turkey

Ercenur Ünal (ercenur.unal@ozyegin.edu.tr)
Özyeğin University, İstanbul, Turkey
Nişantepe Mahallesi, Orman Sokak, 34674, Çekmeköy, İstanbul, Turkey

Abstract

We examine the relation between event segmentation in language and cognition in the domain of motion events, focusing on Turkish, a verb-framed language that segments motion paths in separate linguistic units (verb clauses). We compare motion events that have a path change to those that did not have a path change. In the linguistic task, participants were more likely to use multiple verb phrases when describing events that had a path change compared to those that did not have a path change. In the non-linguistic Dwell Time task, participants viewed self-paced slideshows of still images sampled from the motion event videos in the linguistic task. Dwell times for slides corresponding to path changes were not significantly longer than those for temporally similar slides in the events without a path change. These findings suggest that event units in language may not have strong and stable influences on event segmentation in cognition.

Keywords: event cognition; event segmentation; motion events; dwell time; eye-tracking

Introduction

Our visual experience of the world consists of continuous stream of input that we rapidly organize into discrete event units. Events can be perceived and organized in a hierarchical manner, with coarse units subsuming finer units. This hierarchical organization involves the segmentation of events in different levels of granularity (Zacks et al., 2001). For example, we might construe a trip to a coffee shop as one large event unit or as a series of finer units such as walking into the coffee shop, approaching the counter, ordering the coffee, finding a seat, drinking the coffee, leaving the coffee shop. In addition to perceiving events, we frequently communicate about the events we experience at different levels of specificity. How does the way we segment events in language and cognition relate to each other?

Event Segmentation in Cognition

According to Event Segmentation Theory (Zacks et al., 2007; see also Radvansky & Zacks, 2017) our minds naturally segment continuous experiences into meaningful events, which function as the foundational units for memory and cognition. Working memory representations of events, known as *event models*, maintain information about core features of an event including its participants and spatiotemporal framework. When core features of the event are stable, the event model can make accurate predictions about upcoming happenings. However, when core features of the event change, the event model cannot accurately predict what will happen next and has to be updated. This update is perceived as an event boundary. An event boundary is triggered when there are changes to the people or objects in the event (Zacks et al., 2009), intentionality (Baldwin et al., 2001; Saylor et al., 2007), goal-related structure (Zacks, 2004; Kosie & Baldwin, 2021), causality (Cohen & Oakes, 1993), and spatial characteristics such as direction (Zacks, 2004).

The majority of empirical evidence on event segmentation is based on explicit judgments of event boundaries. In a classical paradigm developed by Newtson (1973), participants are instructed to watch a movie showing an actor performing several activities and to indicate each new unit with a button press. Studies using this paradigm have shown that people can consistently identify the breakpoints in the event in a hierarchical manner (Sargent et al., 2013; Sasmita & Swallow, 2023; Zacks et al., 2001). Furthermore, the movie frames that correspond to the breakpoints are remembered better than the frames that correspond to within-unit moments (Newtson & Engquist, 1976). However, one caveat of this method is that it requires substantial verbal instruction and explanation regarding events and what 'fine' and 'coarse' event units might correspond to.

This caveat is addressed by the Dwell Time paradigm developed by Hard and colleagues (2011). Participants are presented with a series of still images sampled from videos of continuous actions at regular intervals. They are asked to advance through the slideshow at their own pace by pressing a button while the time between the button presses (i.e., how long they dwell on each image) is recorded. Studies using this method have shown that slides that correspond to event boundaries are viewed for a longer time compared to the slides that fall between event boundaries (Hard et al., 2011; Kosie & Baldwin, 2021; Meyer et al., 2011; Sage & Baldwin, 2014; Zheng et al., 2020). This increase in looking time is attributed to the additional demand for attention at event boundaries that is required for consolidating event representations in memory. Importantly, the looking time data from the Dwell Time task correlates with explicit boundary judgments (Hard et al., 2001). Due to its non-verbal and implicit nature this method has been deemed suitable for use with diverse populations, including children (Kosie & Baldwin, 2021; Meyer et al., 2011).

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Event Segmentation in Language

Beyond perception, people frequently communicate about the events they perceive. Each event comprises several components and there are language-specific constraints regarding how much information about event components can be packaged in a single linguistic unit (i.e., a verb phrase; Levelt, 1989).

A well-attested case is motion events (Talmy, 2000; Bohnemeyer et al., 2007; see also Ünal et al., 2023). Motion events (e.g., *a ball rolled into a house*), involve a moving entity known as the figure (a ball), with respect to a landmark known as the ground (a house), along a trajectory or path (into) in a certain way or manner (rolling).

According to Talmy's (2000) typological classification, languages differ based on whether they encode path of motion in verbs or in satellites of the verb. In satellite-framed languages (e.g., English, German, Dutch) manner of motion is typically expressed in the main verb, while path is expressed through satellites (e.g., particles or adpositions). Since manner verbs can be freely combined with different path segments, satellite-framed languages can tightly package manner and path into a single clause (or verb phrase), even if there are multiple path segments (see example 1).

(1)	The ball	rolled	down	into	the house
		verb	preposition	preposition	
		manner	path	path	

By contrast, in verb-framed languages (e.g., French, Turkish, Spanish) path of motion is typically expressed in the main verb, while manner of motion is optionally expressed through adverbs, subordinate verbs or adpositions. As a result, each path segment or change in the direction of motion is expressed in a new verb phrase (see example 2).

(2)	Top	yuvarlan-arak	in-di	ve	ev-e	gir-di
	ball	roll-CONN	descend-PST	and	house-DAT	enter-PST
		sub. verb	verb			verb
		manner	path			path

'The ball rollingly descended and entered the house.'

Linguistic Influences on Event Segmentation

Could cross-linguistic differences in the segmentation of events influence event segmentation in language? This issue relates to a broader discussion the relation between language and other aspects of cognition (Wolff & Holmes, 2010; Ünal & Papafragou, 2016). In one view, semantic distinctions in language create stable differences in how speakers of different languages reason about events even when they are not using language (Levinson, 2003; Majid et al., 2004). On an alternative view, event categories are shared between speakers of different languages to a large extent (Gleitman & Papafragou, 2016; Landau et al., 2010).

Until recently, these competing views could not be tested in the domain of event segmentation as linguistic and non-linguistic segmentation of events have been examined independently. However, a recent study by Gerwien and von Stutterheim (2018) investigated this by comparing native speakers of French (a verb-framed language) and German (a satellite-framed language). The study examined formation of event units in motion events that involve orientation/direction change through both linguistic and non-linguistic tasks. In the linguistic task, participants described motion events shown in brief video clips. French speakers were significantly more likely to produce more than one verb phrase to describe the motion events that had a change in orientation/direction. In contrast, German speakers were less likely to do so. In the non-linguistic task, another group French and German speakers performed the Newtson task (1973) to segment the same motion events. That is, participants were asked to press a button when they perceive a change in the situation in the video. French speakers were more likely to indicate that there was an event boundary compared to German speakers. Specifically, French speakers' detection of the additional event boundary corresponded to the points when the figure changed direction/orientation. The cross-linguistic differences between French and German observed in both linguistic and non-linguistic tasks were taken as evidence for the presence of strong language-driven influences on cognitive event unit formation.

Although the findings reviewed above reveal a parallel between event units in language and cognition, a few aspects of the study challenge this interpretation. First, in the linguistic task, the descriptions were only coded in terms of whether or not the participant used more than one verb phrase referring to the motion events. However, the content of the phrases (path, manner or both) were not considered. This is important because even though in verb-framed languages manner is typically encoded outside of the main verb, manner verbs can also be used (even though there are fewer manner verbs available). Thus, it might be possible to produce a description consisting of more than one verb phrase without using a new phrase to refer to a different path segment but simply by expressing manner and path—two motion components occurring simultaneously—in separate verb phrases (e.g., rolled and entered).

Second, it is possible that in the non-linguistic task participants may have been implicitly verbalizing even though they were not explicitly required to respond verbally. This is especially likely for the study by Gerwien and von Stutterheim (2018) given that participants completed boundary judgments on the Newtson task (1973) by interpreting verbal instructions (i.e., to press a button "when [they] perceive a change/whenever something new happens"). Due to its explicit and verbal nature, this task might have encouraged linguistic encoding during non-linguistic event segmentation, even if the response was non-verbal (i.e., button press). In fact, previous work has shown that cross-linguistic differences disappear when people are prevented from implicitly using language while performing a non-linguistic task (e.g., Trueswell & Papafragou, 2010; Winaver et al., 2007), leaving open the possibility that the cross-linguistic differences reported by Gerwien and von

Stutterheim (2018) reflects thinking with language rather than non-linguistic cognition.

The Current Study

In the present study, we revisit the relation between linguistic and non-linguistic event segmentation by addressing the limitations of prior work discussed above. We focus on native speakers of Turkish, a verb-framed language that allows expressing motion events in multiple clauses with separate verbs for each path segment. We use a within-language comparison by examining event segmentation with linguistic and non-linguistic measures within the same language population to understand the nature of language-cognition interactions in event segmentation.

We address the limitations of previous work in the following ways. First, we conduct a more detailed examination of event descriptions in language. As described in the previous section, people can distribute path and manner into different verb phrases, but since these components often occur simultaneously, expressing them with different phrases in language may not necessarily indicate that they correspond to different cognitive event units. Therefore, we delve into the content of the descriptions to ensure that descriptions with multiple verb phrase indeed convey a change in direction/orientation. Second, we use the Dwell Time paradigm (Hard et al., 2011), which offers a more implicit measure of event segmentation. To further validate this measure, we also record participants' eye movements as an index of attention allocation as they complete the Dwell Time task.

In the linguistic task, participants described videos of motion events that involved a change in direction of motion. As a control, we also included motion events that do not involve a change in direction. Of interest was whether participants would be more likely to produce descriptions that consist of multiple verb phrase for events that involve a path change as opposed to those that do not have a path change. Also, of interest was whether events described using multiple verb phrase were indeed due to the use of two or more path verbs, especially for events that involve a change in direction of motion.

In the non-linguistic task, participants completed an eye-tracked Dwell Time task. We examined whether there would be an increase in Dwell Time for the images that corresponded to the event boundaries (i.e., change in direction of motion). Assuming that our predictions for the linguistic task surface in our data, there are two possibilities for the non-linguistic task. If language shapes how people perceive events, Turkish-speakers should segment events that involve a path change vs. those who do not differently. Thus, Dwell Time should increase for the slides that depict a path change. Alternatively, if event categories are largely shared regardless of language, then number of event units in the segmentation task may not necessarily parallel the number of event units in language.

Method

Participants

Data were collected from native speakers of Turkish (n = 31, 21 females, M_{age} = 21.42 years, SD_{age} = 2.33, range = 18 - 27). Participants were undergraduate students at Özyeğin University and received course credits for their participation. Data from two additional participants were excluded because one was not a native Turkish speaker and the other did not follow the instructions.

Stimuli

Target stimuli consisted of motion events depicting geometric objects changing location with a specific manner and along a trajectory with respect to a landmark object. There were two types of motion events: events that involved a path change (e.g., a ball rolled down a hill, and moved along the lateral axis to enter into a house; Fig.1A) and events that did not involve a path-change (e.g., a ball spun along the lateral axis to enter into a tower; Fig.1B). There were also filler events that did not involve any path change and any landmark (e.g., a triangle jumping to the left).

Stimuli were created in Adobe Premiere Pro CC 2015. Each video was 10-seconds long. Videos featured a sky-blue background with a cloud and green grass on the ground. Each video incorporated a landmark object, combined with a moving figure to establish distinct motion paths. For *into* paths, landmarks were positioned near the motion's end point. For *past* paths, landmarks were placed towards the motion's end in a way that allows the moving object to pass them. Paths involving *upward* or *downward* motion were represented using a hill or stairs.

For the non-linguistic task, we created a slideshow by sampling screenshots from the videos at 1-second intervals. Thus, each 10-second video was converted into a slideshow consisting of 11 slides (see Fig.1). For path-change events, the slide corresponding to the change in direction was set to be in the middle, that is, on the 6th slide, for each slideshow.

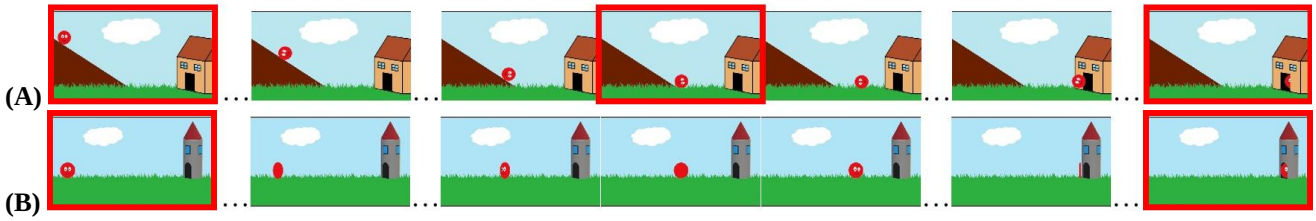

Figure 1: Examples of boundary (outlined in red) and non-boundary slides: (A) path-change (B) no path-change

Norming study Prior to the experiment, event boundaries which corresponded to path changes were independently verified in a norming study. Instructions and procedure were adapted from Kosie and Baldwin (2021) that used a similar Dwell Time task. Participants (n = 30) were shown the complete sequence of slides in order and asked whether each slide depicted a boundary (1) or not (0). Mean boundary judgments for the path-change slide (slide 6) and the slide leading up to the path-change (slide 5) were significantly higher for events that involved a path change (M = 0.46) as opposed to the events that did not involve a path change (M = 0.30, χ^2 (1) = 9.238, p = .002).

Procedure

Participants were tested individually in the lab. They were seated approximately 60cm away from a DELL Precision M4800 laptop with the SMI RED 250 eye-tracker (SensoMotoric Instruments) mounted underneath the screen. The stimuli were presented via NBS Presentation software (Version 23.1, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

All participants received both the linguistic and non-linguistic Dwell Time task in a fixed order: they performed the non-linguistic task first and the linguistic task second. This was done to prevent transfer from a task that involves using language to one that does not.

Linguistic Task Participants watched nine video clips of motion events displayed on the computer. After watching each video clip, they described what happened in the video to a confederate addressee who was seated across them and could not see the computer display. Individual videos were presented in a single randomized order, with half of the participants receiving the original list and the remaining half receiving the items in the reversed order.

Non-linguistic Dwell Time Task Participants were instructed that they would see pictures on the screen and that they could advance through the images at their own pace by pressing the spacebar on the keyboard. They were asked to keep their heads as still as possible so that the eye-tracker could track their eyes accurately. Each participant saw the total set of nine slideshows. Individual slideshows were presented in a single randomized order with half of the participants receiving the original order and the other half receiving the reversed order.

Prior to the main task, participants completed a practice trial that was meant to acquaint them with pressing the spacebar to watch an event unfold. After the practice trials, a 5-point calibration and validation procedure were completed. Following an opportunity to ask questions, the main task began. We recoded the latency between button presses as an index on the total amount of time spent on a given slide as well as the total duration of fixations to the slide when it remained on the screen.

Coding

Descriptions were transcribed and coded by native Turkish speakers on ELAN (Lausberg & Sloetjes, 2009). We coded the number of linguistic units participants used when describing the event. Following Gerwien and von Stutterheim (2018), a linguistic unit was defined as a one finite verb that referred to the motion event shown in the video. Additionally, we examined the semantic features of the descriptions that consisted of multiple verb phrase. Specifically, we coded whether or not descriptions consisting of multiple verb phrases consisted of at least two different path verbs. For example, descend + enter contains multiple path verbs whereas roll + enter contains manner and path distributed into different verbs. Clauses or verb phrases that did not refer to the motion event were not counted.

Preprocessing of the Eye Gaze Data

A message was sent from the Presentation software to the eye-tracker to indicate the onset and offset of each slide. A rectangular Area of Interest (AoI) was defined to cover each slide using SMI BeGaze software, to ensure participants' attention were directed to stimuli. Fixation durations to the AoI were computed by the SMI BeGaze software.

Using R script version 4.2.3 (R Core Team, 2023), we assessed whether participants experienced more than 45% trackloss throughout the entire task or if individual trials had over 50% trackloss. None of the participants exhibited more than 45% trackloss across all trials in the non-linguistic task; consequently, no participant was excluded from the analyses. We excluded trials with more than 50% trackloss (1.26% of all data).

Results

Linguistic Task

The speech data were analyzed using generalized binomial linear mixed effects models. Models were fit using *glmer* function in *lme4* package (version 1.1.33; Bates et al. 2015) in R (version 4.2.3; R Core Team, 2023).

First, we examined the variation in the number of verb phrases used by participants when describing different types of events. The model tested the fixed effect of event type (path-change, no path-change) on the binary dependent variable of using more than one verb phrase (1 = more than one verb phrase, 0 = single verb phrase) at the trial level (see Figure 2). The fixed effect of event type was tested with sum-to-zero contrasts (-0.5, 0.5) (Schad et al., 2020). The model also included random intercepts for Subjects only (Baayen, 2008; Baayen, Davidson, & Bates, 2008). A model that also included random intercepts for Items produced a singular fit error, so this term was omitted. Results revealed a significant fixed effect of event type: participants were more likely to use more than one verb phrase for path-change events compared to no path-change events (β = 1.0079, SE = 0.3594, z = 2.2804, p = .005).

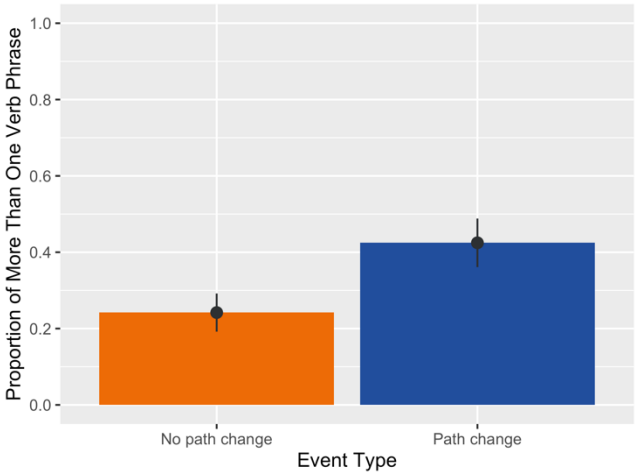

Figure 2: Proportion of more than one verb phrase across event types

Next, we tested how the number of path verbs used in speech differed depending on event type. The model tested the fixed effect of event type (path-change, no path-change) on the binary dependent variable (1 = two or more path verbs, 0 = less than two path verbs) at the trial level (see Figure 3). We used the same strategies for fitting the fixed effects and the random effects structure as in the previous model. Results revealed a significant fixed effect of event type: participants were more likely to use two or more path verbs in descriptions of path-change events compared to no path-change events (β = 1.2317, *SE* = 0.3872, *z* = 3.181, *p* = .001).

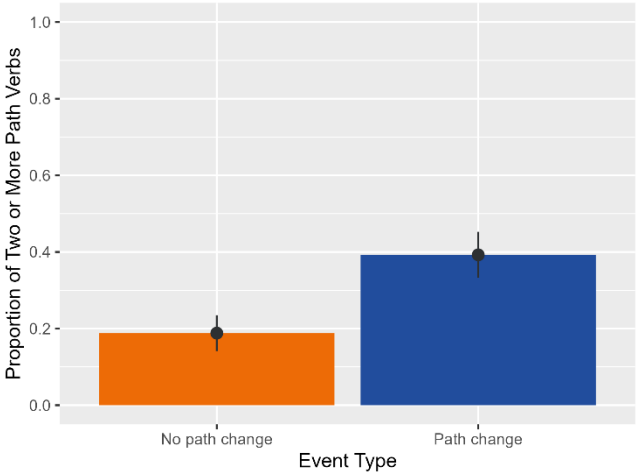

Figure 3: Proportion of descriptions with at least two path verbs across event types

Non-linguistic Dwell Time Task

First, we tested the correlation between the two Dwell Time measures: total duration of fixation during a slide and the latency of the button presses. There was indeed a strong positive correlation between these two measures (*r* = .84).

We began by analyzing fixation durations as they provide a more precise estimate of attention directed at each slide. Following the previous Dwell Time studies (e.g., Hard et al., 2011; Kosie & Baldwin, 2019a, 2019b, 2021) data from the first and last slides from each slideshow were removed prior to analyses. To prepare the Dwell Time data for analyses, we adhered to standard procedures including log transformation and excluding trials that had fixation durations 3 standard deviations above or below the mean (0.57% of the data).

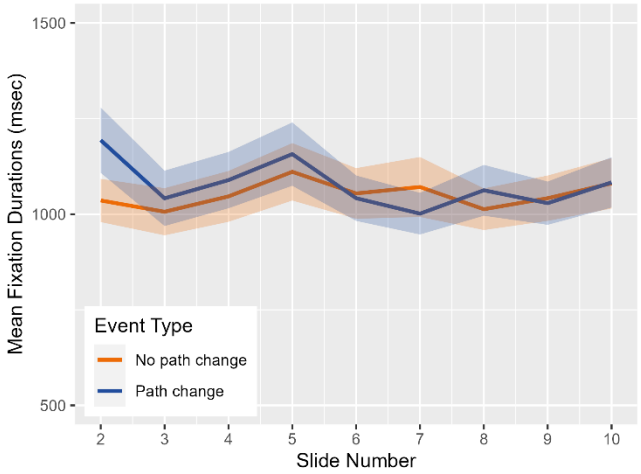

Figure 4: Mean fixation durations across event types and slides

Fixation durations were analyzed with a linear mixed effects model. The model was fit using *lmer* function in *lme4* package (version 1.1.33; Bates et al. 2015) in R (version 4.2.3; R Core Team, 2023). The model tested the fixed effect of event type (path-change, no path-change) on the fixation durations to slides depicting a path change and leading up to the path change as the dependent measure. The fixed effect of event type was tested with sum-to-zero contrasts (-0.5, 0.5) (Schad et al., 2020). The model also included random intercepts for Subjects and Items. The fixed effect of event type was not statistically significant (β = 0.0002591, *SE* = 0.0152, *t* = 0.017, *p* = .987): fixation durations were similar between events that involved a path-change and events that did not involve a path-change (see Figure 4).

To maintain similarity to previous work, we also replicated our analysis using the latency of the button presses as an index of the Dwell Time. We followed the same procedures for data transformation and exclusion, as well as model fitting as described above. This model also did not reveal a significant fixed effect of event type (β = 31.303, *SE* = 70.448, *t* = 0.444, *p* = .674): the latency of the button presses was similar when the slides depicted a path change as opposed to when they did not depict a path change.

Discussion

People readily segment continuous actions that unfold around them into event units and communicate about these events. However, the precise nature of the connection between event

units in language and cognition remains debated. In the present study, we revisit this issue by testing how speakers of Turkish—a verb-framed language that segments motion paths into different verb phrases—describe and perceive motion events.

As a first step, we asked whether participants would be more likely to express a change in direction of motion by using multiple verb phrase for events involving a path change. As expected, our Turkish-speaking participants were more likely to use two or more of verb phrase when describing motion events involving a change in direction of motion compared to the motion events that did not involve a change in direction. A closer inspection of participants' descriptions confirmed that the descriptions consisting of multiple verb phrase indeed had two or more path verbs. Furthermore, such uses were more frequent in the descriptions of events with a path change compared to events without a path change. These findings cohere with previously reported typological patterns in motion event encoding in verb-framed languages (Bohnemeyer et al., 2007; Gerwien & Stutterheim, 2018; Talmy, 2000). They also establish that segmentation of motion paths in language is a good test case for investigating the extent to which language might affect segmentation of motion paths in non-linguistic cognition.

Next, we investigated whether these linguistic patterns would be reflected in how motion events are segmented. Recall that there were two possibilities regarding linguistic influences on non-linguistic event segmentation. According to one possibility, Turkish-speakers were expected to segment events involving a path change differently compared to those that did not. This difference would be reflected as an increase in Dwell Time on slides corresponding to the path change (for events that have one) (Levinson, 2003; Majid et al., 2004). According to an alternative possibility, the number of event units in non-linguistic cognition would not necessarily parallel the number of event units in language (Gleitman & Papafragou, 2016; Landau et al., 2010). In line with the second possibility, our findings revealed that Dwell Times for slides corresponding to path changes for events that involve a path change were not significantly longer than temporally similar slides in events that did not involve a path change, and hence did not depict a change in direction of motion. These data provide evidence against the position that event units in language have stable influences on how people perceive and reason about events.

Our findings seem to diverge from the findings of Gerwien and von Stutterheim (2018) showing that speakers of French and German differ in how they segment motion paths in cognition in line with cross-linguistic differences in motion event encoding in language. We believe these differences can be attributed to the nature of the non-linguistic event segmentation tasks used across the two studies. While the previous study used the Newtson task (1973) that measures explicit judgments of event boundaries, our study relied on the relatively more implicit Dwell Time paradigm (Hard et al., 2011). As discussed earlier, the explicit and verbal nature of the Newtson task might have encouraged participants to rely on language during event segmentation judgments, resulting in discontinuities across speakers of French and German. By contrast, our findings suggest that event units in language do not shape how coarsely or finely event units are segmented in cognition, at least under conditions that do not encourage reliance on language.

Our norming data corroborates this explanation. Despite the fact that Dwell Times did not increase on slides corresponding to path changes, in the norming task participants were more likely to indicate the presence of a boundary for such slides compared to temporally similar slides for events that did not involve a path change. This norming data was elicited with similar (verbal and explicit) instructions as the Newtson task (1973). Note that even though they were statistically significant, the proportions of placing a boundary in both the norming study and the linguistic task are slightly lower than 50%.

Our findings are also consistent with the findings of another cross-linguistic study on event segmentation comparing speakers of Dutch and Avatime (Defina, 2016). Unlike Dutch, Avatime is a language with pervasive use of serial verb constructions, where two or more verbs appear consecutively, forming a compound structure that refers to a single event. However, despite this cross-linguistic difference, speakers of Avatime and Dutch did not differ in a Dwell Time segmentation task, with only an effect related to familiarity with events being observed. In another study, differences in non-linguistic event segmentation emerged when Avatime speakers were primed with serial verb constructions compared to coordinate clauses prior to the Dwell Time task. Together, these findings suggest that linguistic influences on event segmentation—if any—seems to be limited to situations in which language is explicitly or implicitly used during non-linguistic event segmentation.

Finally, our findings converge with another line of work on motion event expressions in speech and gestures that approach to the relation between verbal and non-verbal event segmentation from a different perspective. This line of work has shown that when people describe motion events, each co-speech gesture expresses the semantic information (i.e., path, manner, or both) encoded within a verbal clause in the accompanying speech (Kita & Özyürek, 2003). For example, speakers of Turkish (and other verb-framed languages) use separate gestures to express manner and path. However, when people are asked to describe motion events only using gestures without accompanying speech, speakers of both Turkish and English conflated path and manner into a single gesture (Özçalışkan et al., 2016; 2018). This contrast in the gestures produced with and without accompanying speech is also reflected in our linguistic and non-linguistic tasks: our Turkish-speaking participants perceived multiple motion paths as part of a single event unit even though they distributed this information into multiple units in language. Overall, our results strongly suggest that Turkish speakers can flexibly shift between different units of representations when segmenting events in language and cognition.

Acknowledgments

This work was supported by grant 121K259 awarded by TÜBİTAK to E. Ü. We thank Ceren Barış for assistance with stimuli preparation and Teoman Soydan, Elif Muşlu, Sarp Kut Güler, Dursunay Gibiroğlu, and Eren Koçyiğit for their assistance with data collection and annotation.

References

Baayen, R. H. (2008). Analyzing linguistic data a practical introduction to statistics using R. New York, NY: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*(3), 708–717. https://doi.org/10.1111/1467-8624.00310

Bates, D., M¨achler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bohnemeyer, J., Enfield, N. J., Essegbey, J., Ibarretxe-Antunano, Iraide., Kita, S., Lüpke, Friederike., & Ameka, F. K. (2007). Principles of event segmentation in language: The case of motion events. *Language*, *83*(3), 495–532. https://doi.org/10.1353/lan.2007.0116

Cohen, L. B., & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology*, *29*(3), 421–433. https://doi.org/10.1037/0012-1649.29.3.421

Defina, R. (2016) Events in Language and Thought: The Case of Serial Verb Constructions in Avatime, Radboud University Nijmegen (PhD Thesis)

Gerwien, J., & von Stutterheim, C. (2018). Event segmentation: Cross-linguistic differences in verbal and non-verbal tasks. *Cognition*, *180*, 225–237. https://doi.org/10.1016/j.cognition.2018.07.008

Gleitman, L. R., & Papafragou, A. (2016). New perspectives on language and thought. In K. Holyoak & R.Morrison (Eds.), The Oxford handbook of thinking and reasoning (2nd ed., pp. 543–568). Oxford University Press.

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586–604. https://doi.org/10.1037/a0024310

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kosie, J. E., & Baldwin, D. (2019). Attentional profiles linked to event segmentation are robust to missing information. *Cognitive Research: Principles and Implications*, *4*(1), 1-18. https://doi.org/10.1186/s41235-019-0157-4

Kosie, J. E., & Baldwin, D. (2019). Attention rapidly reorganizes to naturally occurring structure in a novel activity sequence. *Cognition*, *182*, 31-44. https://doi.org/10.1016/j.cognition.2018.09.004

Kosie, J. E., & Baldwin, D. A. (2021). Dwell Times showcase how goal structure informs preschoolers' analysis of unfolding motion patterns. *Child Development*, *92*(6), 2235–2243. https://doi.org/10.1111/cdev.13661

Landau, B., Dessalegn, B., & Goldberg, A. M. (2010). Language and space: Momentary interactions. In P. Chilton & V. Evans (Eds.), Language, cognition, and space: The state of the art and new directions (pp. 51–78). London: Equinox Publishing.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, *41*(3), 841–849. https://doi.org/10.3758/BRM.41.3.841

Levelt, W. (1989). Speaking. MIT Press.

Levinson, S. C. (2003). *Space in language and cognition explorations in cognitive diversity*. Cambridge University Press.

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108-114. https://doi.org/10.1016/j.tics.2004.01.003

Meyer, M., Baldwin, D. A., & Sage, K. (2011). Assessing young children's hierarchical action segmentation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28–38. https://doi.org/10.1037/h0035584

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436-450. https://doi.org/10.1016/0022-1031(76)90076-7

Özçalışkan, Ş., Lucero, C., & Goldin-Meadow, S. (2016). Does language shape silent gesture? *Cognition*, *148*, 10–18. https://doi.org/10.1016/j.cognition.2015.12.001

Özçalışkan, Ş., Lucero, C., & Goldin-Meadow, S. (2018). Blind speakers show language-specific patterns in co-speech gesture but not silent gesture. *Cognitive Science*, *42(3)*, 1001–1014. https://doi.org/10.1111/cogs.12502

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current opinion in behavioral sciences*, *17*, 133-140. https://doi.org/10.1016%2Fj.cobeha.2017.08.006

Sage, K. D., & Baldwin, D. (2014). Looking to the hands: Where we dwell in complex manual sequences. *Visual Cognition*, *22*(8), 1092-1104. https://doi.org/10.1080/13506285.2014.962123

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., ... & Beck, T. M. (2013). Event segmentation ability uniquely predicts event

memory. *Cognition*, *129*(2), 241-255. https://doi.org/10.1016/j.cognition.2013.07.002

Sasmita, K., & Swallow, K. M. (2023). Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods*, *55*(1), 428-447. https://doi.org/10.3758/s13428-022-01832-5

Saylor, M. M., Baldwin, D. A., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development*, *8*(1), 113–128. https://doi.org/10.1207/s15327647jcd0801_6

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of memory and language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Talmy, L. (2000). A typology of event integration. In Toward a cognitive semantics: Typology and process in concept structuring (Vol. 2, pp. 213–288). MIT Press.

Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language*, *63*(1), 64–82. https://doi.org/10.1016/j.jml.2010.02.006

Ünal, E., Mamus, E., Özyürek, A. (2023). Multimodal encoding of motion events in speech, gesture and cognition. *Language and Cognition*, 1-20. https://doi.org/10.1017/langcog.2023.61

Ünal, E., & Papafragou, A. (2016). Interactions between language and mental representations. Language Learning, 66, 554-580. https://doi.org/10.1111/lang.12188

Winaver, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, *104*(19), 7780-7785. https://doi.org/10.1073/pnas.0701644104

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 253-265. https://doi.org/10.1002/wcs.104

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive science*, *28*(6), 979-1008. https://doi.org/10.1207/s15516709cog2806_5

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29-58. https://psycnet.apa.org/doi/10.1037/0096-3445.130.1.29

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, *138*(2), 307–327. https://doi.org/10.1037/a0015305

Zheng, Y., Zacks, J. M., & Markson, L. (2020). The development of event perception and memory. *Cognitive Development*, *54*, 100848. https://doi.org/10.1016/j.cogdev.2020.100848