Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification

PIA KNOEFERLE,[a] THOMAS P. URBACH,[b,c] AND MARTA KUTAS[b,c,d,e]

[a]Cognitive Interaction Technology Excellence Cluster, Bielefeld University, Bielefeld, Germany
[b]Department of Cognitive Sciences, University of California San Diego, La Jolla, California
[c]Center for Research in Language, University of California San Diego, La Jolla, California
[d]Department of Neurosciences, University of California San Diego, La Jolla, California
[e]Kavli Institute for Brain and Mind, University of California San Diego, La Jolla, California

Abstract

To re-establish picture-sentence verification—discredited possibly for its over-reliance on post-sentence response time (RT) measures—as a task for situated comprehension, we collected event-related brain potentials (ERPs) as participants read a subject-verb-object sentence, and RTs indicating whether or not the verb matched a previously depicted action. For mismatches (vs. matches), speeded RTs were longer, verb N400s over centro-parietal scalp larger, and ERPs to the object noun more negative. RTs (congruence effect) correlated inversely with the centro-parietal verb N400s, and positively with the object ERP congruence effects. Verb N400s, object ERPs, and verbal working memory scores predicted more variance in RT effects (50%) than N400s alone. Thus, (1) verification processing is not all post-sentence; (2) simple priming cannot account for these results; and (3) verification tasks can inform studies of situated comprehension.

Descriptors: Picture-sentence verification, Situated comprehension, Event-related brain potentials

Although humans can readily understand sentences about events that are displaced in space and time without being present at the scene, language users often are physically "situated" in the scene. Indeed, information from a co-present or recently experienced visual environment has been found to affect auditory sentence comprehension with a relatively short lag. This has been evidenced by the continuous monitoring of eye gaze in a visual context (act-out tasks: e.g., Chambers, Tanenhaus, & Magnuson, 2004; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; passive listening comprehension: e.g., Altmann, 2004; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Knoeferle & Crocker, 2007), and by event-related brain potentials (ERP) recordings during concurrent utterance comprehension and visual scene inspection (e.g., Knoeferle, Habets, Crocker, & Münte, 2008). Processing accounts of situated utterance comprehension accommodate this temporally coordinated interplay: e.g., language directs attention to objects and events (or representations thereof in visuo-spatial working memory) and can, in turn, receive rapid feedback from scene-based mental representations (Knoeferle & Crocker, 2006; 2007; Mayberry, Crocker, & Knoeferle, 2009, 'Coordinated Interplay Account').

The focus in the field of psycholinguistics to date has been on the "facilitative" effects of visual contexts (scenes) on language processing, e.g., how visual context might incrementally disambiguate temporary linguistic and/or referential ambiguities (e.g., Altmann, 2004; Chambers et al., 2004; Knoeferle et al., 2005; Knoeferle & Crocker, 2007; Sedivy et al., 1999; Spivey et al., 2002). However, visual context apparently can have effects on language processing even when the accompanying language and/or reference is unambiguous (e.g., Stroop, 1935). Thus, it is important to examine the nature and timing of visual context and language comprehension interactions more generally, even when the two information sources are not completely in accord, which, upon careful consideration, is often the case. Inconsistencies can range from outright mismatches (e.g., when a passenger alerts the driver to a red traffic light, which has turned green by the time they actually look), mismatches due to lexically different but semantically similar terms, to subtle nuances in how different individuals describe the same object or event (e.g., one person sees and thinks "couch" which another might refer to as "sofa" or "a place to relax"). Socio-economic factors (e.g., age, gender,

social status) also may influence not only how people perceive their world but also how they talk about it. An incident that a witness describes as "slightly prodding a friend" may be expressed as "aggressively shoving a man" by another individual. In short, utterances and written text or signs may often be a less-than-perfect match for a language user's current representation of the non-linguistic visual environment. Despite its obvious import for theories of sentence comprehension and situated processing accounts, however, little is known about how inferences about the time course of incremental comprehension might extend to situations when visual context clashes with the linguistic input. Morever, such findings would need to be theoretically accommodated, if observed (e.g., Knoeferle & Crocker, 2006, 2007).

Models of how people understand and evaluate the truth or falsity of sentences against visual contexts have generally proposed a serial comparator mechanism between corresponding picture and sentence constituents, e.g., between dots and the noun "dots" (Carpenter & Just, 1975). These types of models aim to account for the general finding that verification times tend to be shorter when visually depicted information and linguistically expressed information match (vs. not). At first blush, then, verifying the truth of a sentence against a picture might appear well suited for developing cognitive models of language comprehension. In light of the burgeoning interest in situated and embodied language comprehension, the verification paradigm also would seem to be a fruitful approach for investigating the time course of language comprehension in visual contexts. Its utility and the validity of the proposed serial comparison mechanism have, however, been sharply criticized (e.g., Tanenhaus, Carroll, & Bever, 1976; see Carpenter & Just, 1976, for a reply).

In particular, it has been argued that the key dependent measure (i.e., response verification time) may reflect scene-sentence comparison processes that take place only after sentence comprehension has been completed rather than the effects of pictorial information on ongoing language comprehension processes: "At best, such studies may provide specific examples of how subjects can verify sentences they have already understood against pictures they have already encoded" (Tanenhaus et al., 1976). In addition, a few failures to find the typical congruence effect (longer response latencies for a picture-sentence mismatch than match) with serial picture-sentence presentation have led to concerns about the generality of the paradigm. Despite its occasional use to study language comprehension (e.g., Goolkasian, 1996; Reichle, Carpenter, & Just, 2000; Singer, 2006; Underwood, Jebbett, & Roberts, 2004), then, the fact remains that insights obtained with this paradigm have had minimal impact on psycholinguistic theories of online sentence comprehension or of situated sentence comprehension.

Perhaps the conclusions by Tanenhaus and colleagues are justified. After all, punctate measures such as verification response times and accuracy scores obtained at the end of sentences may merely reflect post-sentence verification processes and say nothing about online incremental language processing. Alternatively, however, they may—to an important extent—reflect (visual context effects on) at least some aspects of online language processing—a possibility that can be better assessed if the offline verification time measures are combined with a continuous online processing measure such as ERPs (see, e.g., Kounios & Holcomb, 1992).

The study of language processing benefits from combining these two measures (verification response times [RTs] and sentence ERPs) to the extent that ERPs in verification tasks reflect comprehension and/or verification processes rather than mere mismatch detection (since, in that case, "sentence or word processing" ERPs can be related to "verification" RTs, operationally defined). Critically, manipulations of semantic context that modulate sentence processing are known to be associated with a greater negativity (N400) starting within 200 ms of stimulus onset for semantically anomalous or unexpected words relative to semantically congruous, expected words (Kutas & Hillyard, 1980). Moreover, ample data show that such N400 effects are related to comprehension rather than strictly mismatch or pure verification processes (Fischler, Bloom, Childers, Roucos, & Perry, 1983), do not necessarily index plausibility (Federmeier & Kutas, 1999), and are dissociable from RT effects (Fischler et al., 1983).

Existing ERP evidence indicates that picture-sentence congruence effects can appear incrementally during sentence processing in the form of N400 effects and, sometimes, P600 effects (e.g., Vissers, Kolk, Van de Meerendonk, & Chwilla 2008; Wassenaar & Hagoort, 2007), among other possibilities. For example, some picture-sentence congruence manipulations during sentence comprehension were found to elicit a negativity akin to an N2b component previously observed in response to adjectival color mismatches (object: red square; linguistic input: green square; token test) and taken to index attentional detection of a mismatch rather than language processing (D'Arcy & Connolly, 1999; Vissers et al., 2008). In other cases, however, picture-sentence congruence manipulations yielded N400s effects that resemble those in language comprehension tasks (e.g., auditory N400: Holcomb & Neville, 1991; McCallum, Farmer, & Pocock, 1984), consistent with the view that visual contexts modulate language comprehension processes (e.g., Wassenaar & Hagoort, 2007). Szűcs, Soltész, Czigler, and Csépe (2007) reported both N2b and N400 activity; specifically, they observed N2bs to mismatches when the second of a pair of items (letter, digit) mismatched on color or category membership, while only category mismatches also elicited a centro-parietal N400. Clearly, there is much to be learned about the nature of the relationship between verification and comprehension, and language processing more generally.

The present study thus examines whether, and if so how, participants' picture-sentence congruence effects during sentence processing (that resemble ERP effects in strictly language comprehension tasks) relate to their end-of-sentence response latency verification times. On each trial, a few-second inspection of a clip art scene (e.g., a gymnast punching a journalist) was followed by word-by-word presentation of a written sentence; the verb either matched (e.g., *punches*) the depicted agent-action-patient event or not (e.g., *applauds*). Our primary dependent measures were the ERPs elicited by individual words and the post-sentence verification response times, traditionally taken to reflect the ease or difficulty of processing the match or mismatch between a scene and sentence in a verification task (e.g., Clark & Chase, 1972; Gough, 1965).

Consistent with the literature at large, we expect mismatches to be verified more slowly than matches. With evidence in hand that processing of congruous versus incongruous sentence-picture pairs differs as expected, we will examine whether the ERP manifestations of these processing differences during the sentence resemble canonical ERP congruence effects in sentence comprehension tasks (e.g., Kutas & Hillyard, 1984; Van Berkum, Hagoort, & Brown, 1999). Based on two published reports (e.g.,

Szűcs et al., 2007; Wassenaar & Hagoort, 2007), we do expect to see a centro-parietal N400 effect (greater negativity to the mismatch) at the earliest possible mismatch point between the picture and sentence—namely, the verb. One open question is whether or not there will be any other signs of the picture-sentence discord during the course of sentence processing.

Verb-action mismatch ERP effects could be quite local (restricted to the verb and the response). This would be expected on, for example, a 'priming' account according to which a representation of the depicted action primes the processing of the matching (relative to mismatching) verb as reflected in ERPs (see Barrett & Rugg, 1990; Kutas & Van Petten, 1988; Meyer & Schvaneveldt, 1971). Perhaps because there are no other discrepancies between the depicted elements and the sentence constituents, no other ERP congruence effects would be obtained. However, since participants were asked to indicate whether the sentence was true or false with regard to the immediately preceding picture, we would also expect to see a congruency effect in the post-sentence verification times. On a straightforward priming account, ERP and RT congruence effects would both reflect a similar (priming) process, and thus should co-vary directly (i.e., a large ERP congruence effect should be associated with a large verification time congruence effect and vice versa).

Alternatively, there may be multiple online effects of visual context on sentence processing, which may be manifest with variable and/or complex relations to the offline verification time effects. Referential verb-action mismatches are a good test case for adjudicating between these alternative accounts, since the locus of the mismatch is restricted to the verb. Finding more than one ERP congruence effect during the course of these sentences and/or the absence of direct variation between ERP picture-sentence congruence effects and RT verification effects, for example, would undermine any simple priming account. Moreover, to the extent that the collection of online ERP congruity effects as a whole were able to account for significant variance in the offline verification times, we could infer that verification was not simply something that happened after sentence comprehension was complete. In addition, to facilitate interpretation of any observed

congruence effects, we examined whether, and if so how, the nature or time course of visual context effects on online sentence comprehension was modulated by participants' verbal working memory and/or visuo-spatial abilities. To that end, we obtained visual (Extended Complex Figure Test: Motor-independent version; Fasteneau, 2003) and verbal (Daneman & Carpenter, 1980, reading span test) working memory scores for our participants.

Methods

Participants

Twenty-four students of the University of California San Diego (UCSD) (11 men, 13 women, aged 19–24, mean = 20.04 years) were paid for participation. All participants were native English speakers, right-handed according to the Edinburgh Handedness Inventory, and had normal or corrected-to-normal vision. All gave informed consent in accordance with the Declaration of Helsinki; the experiment protocol was approved by the UCSD Institutional Review Board.

Materials

An example image pair for one item shows a gymnast punching a journalist (Figure 1a) and a gymnast applauding a journalist (Figure 1b). These were paired with one of the following sentences:

(1a) The gymnast punches the journalist.
(1b) The gymnast applauds the journalist.

The experiment was a 1-factor (congruence) within-subjects design with two levels (incongruent versus congruent verb-action relation): Sentence (1a) *The gymnast punches the journalist* is congruent when following Figure 1a and incongruent following Figure 1b. The materials were counterbalanced to ensure that any congruency-based ERP differences were not spuriously due to stimuli or to their presentation: (1) Each verb (e.g., punches/ applauds) and corresponding action (punching/applauding)
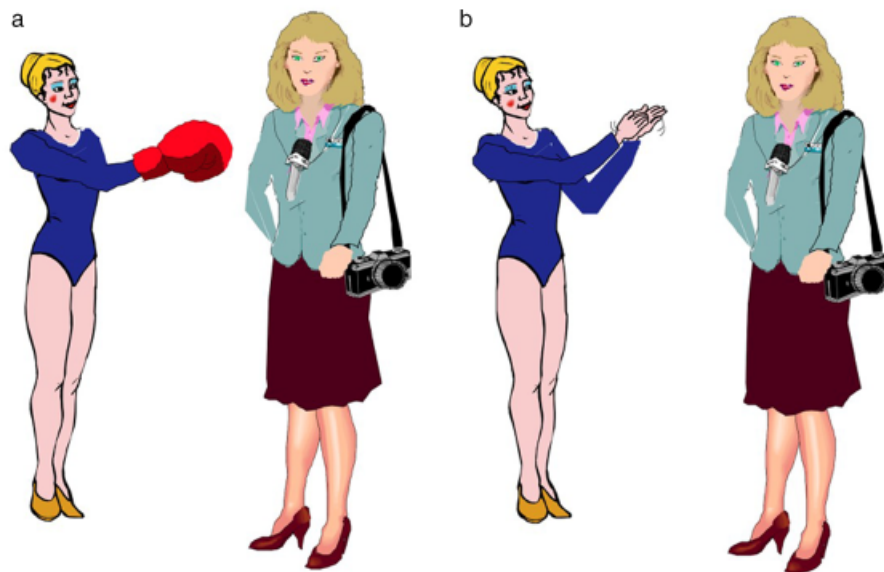

Figure 1. Example image set for the item sentences in (1a) and (1b).

occurred once in a congruent and once in an incongruent condition; (2) Each verb and action occurred in two different items (with different first and second nouns); and (3) Directionality of the actions (the agent standing on the left vs. the agent standing on the right) also were counterbalanced.

We conducted two pre-tests on 168 image and sentence stimuli to determine whether the characters and depicted actions were recognizable as well as the degree of sentence-picture correspondence. Twenty participants (different from those in the ERP study) named the characters and action in each of the 168 scenes. The average accuracy in naming the agent, action, and patient was 81.84%, 85.17%, and 88.75%, respectively. Mean naming accuracy was above 50% for all items. Participants also rated the extent to which each image matched its corresponding sentence on a 5-point scale (1 indicating a poor match and 5 indicating very good congruence). Mean ratings of 4.31 (*SD* 1.0) in the congruent condition, and 2.04 (*SD* 1.14) in the incongruent condition revealed that the congruence manipulation was effective ($p < .001$). Only images for which participants consistently named the characters and the action, and sentences for which there was a greater than or equal to 1.4-point difference between the congruent and incongruent pairs, were selected for the ERP recording.

On the basis of the pretest, we selected 160 images and sentences and constructed 80 item sets, each consisting of 2 sentences and 2 images (such as Figure 1a and 1b, and sentences (1a) and (1b)), which, combined with the within-subject counterbalancing, yielded 8 experimental lists. Each participant saw an equal number of matching and mismatching trials, only one occurrence of an item (i.e., sentence/image pairing), and an equal number of left-to-right and right-to-left action depictions. Assignment of item-condition combinations to a list followed a Latin Square. As a result of the counterbalancing (each verb and action appeared in two different items), cross-modal repetition occurred within half of the items in a list: in other words, a given verb (e.g., "punches") occurred in one item sentence (together with, e.g., a mismatching applauding action), and a depicted punching action occurred in another item (together with a mismatching verb "applauds"). We think this repetition was not noticed since it involved a relatively small number (40) of the overall (240) trials. When asked during the pen-and-paper debriefing whether anything caught their attention ("patterns, anything strange, or surprising"), participants did not report noticing verb-action repetition.

Each list also contained 160 filler items, of which half included one of several different types of mismatches: 'full mismatch' (scene and sentence were entirely unrelated, ensuring that sentence comprehension was not always contingent on the scene), noun-object reference mismatches, mismatches of the spatial layout of the scene, and mismatches of color adjectives. These filler sentences had a variety of different syntactic structures including negation, clause-level and noun-phrase coordination, as well as locally ambiguous reduced relative clause constructions. They included "combinatory" mismatches (i.e., requiring combinatory processing to determine sentence truth). Accordingly, we were relatively confident that participants had to perform combinatory sentence comprehension in the verification task.

Procedure

Each trial consisted of an image (scene) followed by a sentence presented one word at a time in central vision. Images were presented on a 21-inch CRT monitor at a viewing distance of approximately 150 cm. Participants first inspected the image for a minimum of 3000 s, terminated by the participant via right thumb button press. Next, a fixation dot was presented for a random duration between 500 and 1000 ms, followed by a sentence one word at a time, each presented for 200 ms duration with a word onset asynchrony of 500 ms. Note that in picture-sentence verification research, the sentence is often presented as a whole. Rapid serial visual presentation was chosen based on its successful use in ERP-verification studies. Participants were encouraged to focus on comprehension during inspection of both picture and sentence: They were instructed to attend to the picture in order to understand what was depicted, and to comprehend the sentences in the context of the preceding image. To assess comprehension, participants were also required to indicate via a button press as quickly and accurately as possible after each sentence whether it matched ("true") or did not match ("false") the preceding image (only post-sentential responses were "counted" as accurate). Allocation of response hand to "true/false" responses was counterbalanced across lists. Once a participant had responded, there was a short inter-trial interval varying between 500 and 1000 ms, to break the rhythm of the fixed rate of sentence presentation.

In addition to response latencies, we assessed each participant's verbal working memory (VWM) and visual-spatial abilities from a standard working memory test (Daneman & Carpenter, 1980), and the brief, motor-independent version of the extended complex figure test (ECFT-MI, Fasteneau, 2003), respectively. The ECFT-MI recognition trials provide a measure of visual-spatial encoding and recognition processes, and its matching trials verify that a participant's visual-perceptual functions are intact.

ERP Recording and Analysis

ERPs were recorded from 26 electrodes embedded in an elastic cap (arrayed in a laterally symmetric pattern of geodesic triangles approximately 6 cm on a side and originating at the intersection of the inter-aural and nasion-inion lines as illustrated in Figure 2) plus 5 additional electrodes referenced online to the left mastoid, amplified with a bandpass filter from 0.016 to 100 Hz, and sampled at 250 Hz. Recordings were re-referenced offline to the average of the activity at the left and right mastoid. Eye movement artifacts and blinks were monitored via the horizontal (through two electrodes at the outer canthus of each eye) and vertical (through two electrodes just below each eye) electrooculogram. Only trials with a correct response were included in the analyses. All trials were scanned offline for artefacts, and contaminated trials were excluded from further analyses. For 18 of the participants, blinks were corrected with an adaptive spatial filter (Dale, 1994). After blink correction, no more than 25% of the correct trials per condition for a given participant were rejected. All analyses (unless otherwise stated) were conducted relative to a 500-ms pre-stimulus baseline. Of the 80 items, 21 had first and/or second noun phrases that were composite (e.g., 'the volleyball player') while the remaining 59 items had simple noun phrases (e.g., 'the gymnast'). This means that for the 21 items, the first and second nouns have one extra region, rendering analyses and reporting somewhat complicated (i.e., it is unclear which noun region 'volleyball' or 'player' or both to report). Furthermore, as results did not differ substantially as a function of whether analyses were performed across all 80 items or only for the homogeneous subset of 59 items, below we report those for the latter.
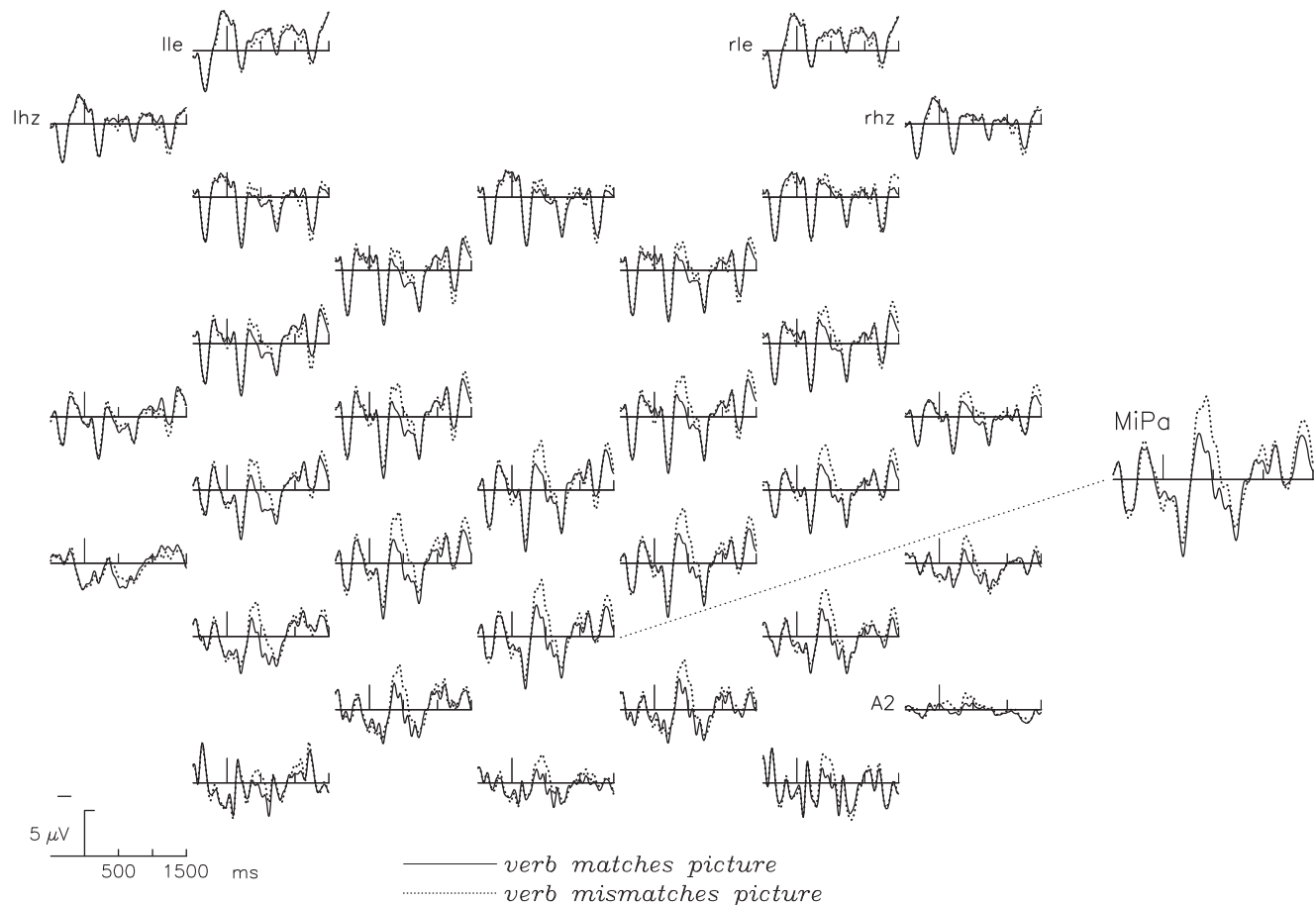
Figure 2. Grand average ERPs (mean amplitude) for all 26 electrodes, right-lateral, left-lateral, right-horizontal, left-horizontal ('rle,' 'lle,' 'rhz,' and 'lhz'), and the mastoid ('A2') at the verb position. Waveforms were subjected to a digital low-pass filter (10 Hz) for visualization. A clear negativity emerges for incongruent relative to congruent sentences at the verb when the mismatch between verb and action becomes apparent. The ERP comparison at the mid-parietal ('MiPa') site is shown enlarged.

Analyses of Behavioral Data

For response latency analyses, any individual score ± 2 standard deviation from the mean response latency of a participant was removed prior to further analyses. Mean response latencies time-locked to the sentence-final word (the second noun) were summarized by participants (*t1*) and items (*t2*) and analyzed via paired sample *t*-tests with the congruence factor (mismatch vs. match). For the analysis of working memory scores from the Daneman and Carpenter (1980) working memory test, we computed the proportion of items for which a given participant recalled all the elements correctly (Conway et al., 2005) as a proxy for VWM scores. For the extended complex figure test, we followed the scoring procedure for the motor-independent ECFT-MI described in Fasteneau (2003).

Analyses of ERP Data

Analyses of variance (ANOVAs) were conducted on the mean amplitudes of the average ERPs elicited by the verb (e.g., *punches*) since this is where the verb-action (in)congruence first becomes evident. We measured the N400 between 300–500 ms post verb onset, as is common for visual N400s, and also from 300–600 ms, since it became apparent that the N400 lasted circa 100 ms into the post-verbal determiner. Based on visual inspection and traditional sensory evoked potential epochs, we also analyzed three earlier time windows (0–100 ms; 100–300 ms;

180–220 ms to explore the P2) in the ERPs to the verb. In addition, we analyzed ERPs elicited by the first nouns (*gymnast*) where we expected to see no effect of congruence and second determiners (*the*) and nouns (*journalist*) where a post-verb effect of congruence was possible in the same time windows. For the second determiner, we analyzed ERPs from 100–300 and 300–500 ms with a baseline from − 500 to − 300 ms to ensure ERP effects at the verb didn't modulate the baseline. The first 100 ms of the second determiner (0–100 ms) were analyzed as part of the verb ERPs (300–600 ms) since visual inspection showed a continuation of verb ERP effects into the early second determiner. We also analyzed the ERP after the offset of the second noun (500–1000 ms; 1000–1500 ms time-locked to the second noun) to gain insight into processing of the verb-action mismatch from sentence end up to the verification response.

We performed omnibus repeated measures ANOVAs on mean ERP amplitudes (averaged by participants for each condition at each electrode site) with congruence (incongruent vs. congruent), hemisphere (left vs. right electrodes), laterality (lateral vs. medial), and anteriority (5 levels) as factors. Interactions were followed up with separate ANOVAs for left lateral (LLPf, LLFr, LLTe, LDPa, LLOc), left medial (LMPf, LDFr, LMFr, LMCe, LMOc), right lateral (RLPf, RLFr, RLTe, RDPa, RLOc), and right medial (RMPf, RDFr, RMFr, RMCe, RMOc) electrode sets (henceforth 'slice') that included con-

gruence (match vs. mismatch) and anteriority (5 levels). Greenhouse-Geisser adjustments to degrees of freedom were applied to correct for violation of the assumption of sphericity. We report the original degrees of freedom in conjunction with the corrected p-values. For pair-wise t-tests that were also conducted to follow up complex interactions, we report p-values after Bonferroni adjustments for multiple comparisons (20 comparisons) unless otherwise stated.

Linear Regression Analyses

We investigated the relationship between ERPs and verification response times. To that end, we conducted correlation, simple, and multiple linear regression analyses to see to what extent end-of-sentence verification times can be predicted from verb ERPs alone, or in combination with subsequent ERP congruence effects. For these analyses, we computed each participant's mean congruence effects (incongruent minus congruent ERP amplitude), and each participant's congruence effect for verification response latencies (incongruent minus congruent latencies). The analyses are thus between two sets of difference scores. Difference scores have been much discussed (and sometimes criticized, e.g., Cohen & Cohen, 1983; Griffin, Murray, & Gonzalez, 1999; Johns, 1981). We think they are informative in our study: the difference scores provide a measure of the extent to which the processing of congruous and incongruous trials differs both at the verb and at the end of the sentence. For response latencies, a positive number indicates longer verification times for incongruous than congruous times and a negative number indicates the converse. For ERPs, a negative number means that incongruous trials were relatively more negative (or less positive) than congruous trials, with the absolute value of the negative number indicating the size of the difference. For the correlation and regression analyses, we relied upon congruence ERP difference scores summarized using the mean across the electrode sites in the four slices used for the ANOVA analyses (e.g., left lateral: LLPf, LLFr, LLTe, LDPa, LLOc; Bonferroni adjustment for 4 analyses in a given window, 0.05/4).

Expectations of which ERP scores should be good predictors in the regression model were influenced by the presence and scalp topography of congruence effects in the ANOVA, the presence of reliable correlations with the verification times, and by published findings showing larger and more robust N400 effects over right (than left) hemisphere sites (see Federmeier & Kutas, 2002). While linear regression is typically used to predict a dependent measure from an independently manipulated variable, we justify its current application (prediction of a dependent variable from another dependent variable) on the view that response latencies result from preceding brain activity.

We screened scatter plots of residuals (errors) against predicted values of the dependent variable to ensure that assumptions of normality, linearity, and homoscedasticity between predicted scores of the dependent variable and errors of prediction were met. There were no multivariate outliers (Mahalanobis distance). We also inspected the data for multicollinearity of the independent variables in the multiple regression models (see Tabachnik & Fidell, 2007); multicollinearity was not an issue for the independent variables in our linear regression models (ERPs from 300–500 ms at the verb and second noun). For multiple regressions, we report adjusted R^2s to ensure that increasing the number of independent variables does not artificially (because of chance variation of the independent variables) increase the accounted-for variance in the response verification congruence

effect. Last but not least, to examine whether there was any reliable difference between two models (e.g., a model with ERP congruence scores at the verb and second noun vs. a nested model with ERP congruence effects at the verb), we computed a partial F (see Hutcheson & Sofroniou, 1999).

Results

Behavioral

Participants' verification accuracy was relatively good (83.84%) overall, suggesting participants understood the sentences and images and their interrelationships, albeit better for congruous (86.93%) than incongruous (80.75%) trials ($t1(1,23) = 1.41$, $p = .17$; $t2(1,58) = 2.36$, $p < .05$). These accuracy values are somewhat lower than those in prior research (e.g., Vissers et al., 2008; Wassenaar & Hagoort, 2007) as we, unlike these others, included a range of picture-sentence filler combinations with reduced relative clause structures, and negated sentences. It is likely that the inclusion of these fillers made it more difficult for participants to verify the picture-sentence mismatch, and thus led to longer response times and lower accuracy levels, than if we had tested exclusively verb-action mismatches. RT analyses also revealed reliably faster verification times for congruent (by participants: 1104 ms, $SD = 271.72$) than incongruent (by participants: 1273 ms, $SD = 361.06$) sentences ($t1(1,23) = -3.14$, $p < .01$, $t2(1,58) = -4.125$, $p < .001$). Participants' average VWM score in the Daneman and Carpenter working memory test was 0.31 (the score reflects the proportion of test elements for which all of the items were recalled correctly; range 0.13–0.65, see Conway et al., 2005). Their average score in the extended complex figure test was 12.13 out of 18 (range 7–18) for the recognition and 9.71 out of 10 for the matching part (range 8–10).

ERPs: Analyses of Mean Amplitudes

Figure 2 displays the grand average ERPs ($N = 24$) at all 26 electrode sites in the picture-congruent versus incongruent conditions time-locked to verb onset. Figure 3 shows the grand average ERPs at prefrontal, parietal, temporal, and occipital sites together with the spline-interpolated topographies of the mean ERP amplitude difference (mismatch minus match) both between 300–500 ms post-verb onset, lasting into the post-verbal determiner, and between 300–500 ms at the second noun (patient). These figures illustrate a broadly distributed negativity (N400) at the verb, maximal at posterior recording sites, and a right lateral negativity at the second noun (patient), both of which are reliably larger for picture-incongruous than picture-congruous verbs. Tables 1 and 2 present the corresponding ANOVA results for the main effect of congruence and interactions between congruence, hemisphere, laterality, and anteriority at the verb and second noun. ANOVAs of ERPs to the first noun (e.g., *gymnast*) and the second determiner (100–300 ms, 300–500 ms) revealed neither reliable main effects nor interactions of picture-sentence congruence (all Fs < 2.53).

Verb

For the verb N400, ANOVAs indicated larger congruence effects over medial (left medial sites: partial eta squared $\eta^2 = 0.50$; right medial sites: $\eta^2 = 0.43$) than lateral (left: $\eta^2 = 0.38$; right: $\eta^2 = 0.42$, for corresponding F-values see Table 1) sites, confirming the typical N400 distribution, maximal over centro-parietal sites. Congruence effects also interacted with anteriority, being smaller over anterior (e.g., LMPf, $\eta^2 = 0.29$; RMPf, $\eta^2 = 0.27$,
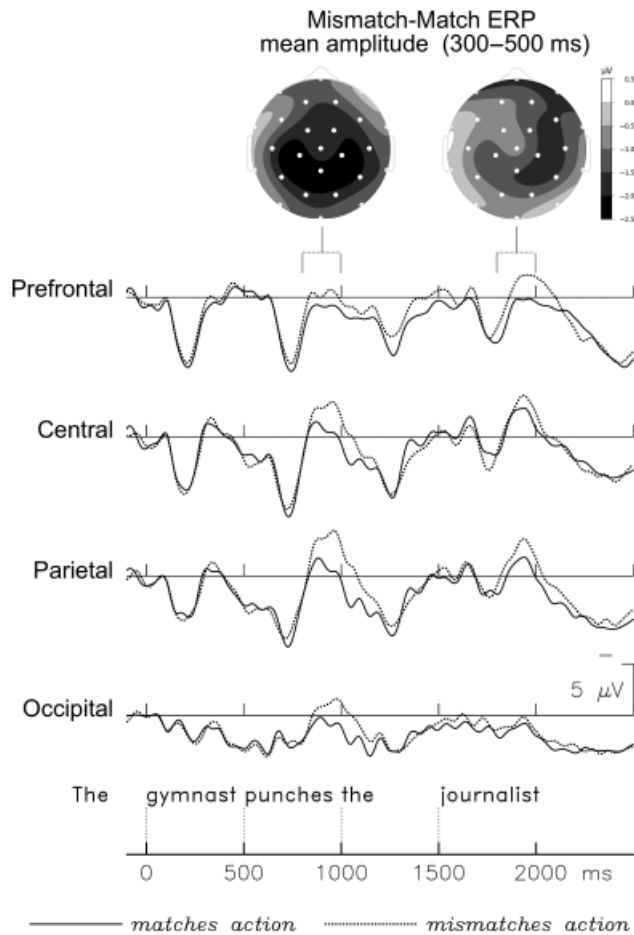
Mismatch-Match ERP mean amplitude (300–500 ms)

Figure 3. Grand average ERPs (mean amplitude) across the sentence at prefrontal, parietal, temporal, and occipital sites. A clear negativity emerges at the verb when the mismatch between verb and action becomes apparent and, later, at the second noun (patient). Spline-interpolated difference scores at the verb and the second noun illustrate the centro-parietal (verb) and right lateral (second noun) distribution of these two ERP effects.

congruence effect n.s.) than central sites (LMCe, $\eta^2 = 0.56$; RMCe, $\eta^2 = 0.46$; $ps < .05$). Complex interactions between congruence, hemisphere, and laterality are brought about by a large difference between congruence effects over left lateral ($\eta^2 = 0.38$) versus left medial ($\eta^2 = 0.50$, difference = 0.12, see Table 1) sites relative to similar effect sizes for lateral ($\eta^2 = 0.42$) and medial ($\eta^2 = 0.43$, difference = 0.01) sites over the right hemisphere. Additional ANOVAs on mean amplitudes in earlier time windows (0–100; 100–300) at the verb revealed no other reliable effects of congruence. Analyses from 300–600 ms at the verb revealed the same pattern of main effects and interactions as from 300–500 ms and are thus not reported in greater detail.

Second Noun ERPs

ERPs for the second noun revealed a reliable interaction between congruence and anteriority from 0–100 ms (Figures 2 and 3), which follow-up analyses showed were reliable over left (but not right) lateral or medial sites (Table 2). For the 300–500 ms (N400) time window, analyses revealed a reliable congruence by hemisphere interaction: None of the main effects of congruence in separate ANOVAS for left/right lateral and medial electrode sites were reliable (Table 2), but effect sizes were still larger over right (lateral $\eta^2 = 0.1$; medial $\eta^2 = 0.06$) than left (lateral $\eta^2 = 0.00$; medial: $\eta^2 = 0.01$) hemispheric sites.

Correlation Analyses

Analyses showed reliable correlations of the response time congruence effect with mean ERP congruence effects summarized over right lateral, medial, and left medial slices at the verb: the larger a participant's verb N400 congruence effect, the smaller her/his response time congruence effect (Figure 5). These analyses with ERPs summarized over the four slices were corroborated at individual electrode sites where we found reliable correlations of the response time verification congruence effects with verb N400 (300–500 ms) congruence effects at RMCe, RDPa, RMOc, RLTe according to Pearson r ($r = .58$, $r = .59$, $r = .59$, $r = .66$, respectively, at these sites, see Figure 4). The only other behavioral measure correlated with (right medial) ERP congruence effects at the verb was verification accuracy: the

Table 1. *Statistical Results Mean Amplitude of ERPs at the Verb*

Sentence position	Time window	Factors	Overall ANOVA	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
Verb	0–100		—				
	100–300		—				
	180–220		—				
	300–500	C	21.85***	13.96**	23.26***	16.70***	17.26***
		CH	—				
		CL	12.99**				
		CA	4.29*	3.89*	5.39*	3.79*	2.26
		CHL	6.61*				
		CHA	3.73*				
		CLA	—				
		CHLA	3.51*				

Note: Columns 3–4 show the results of the overall ANOVA electrode sets at the verb (20 electrode sites); columns 5–8 show results of separate follow-up ANOVAS for left lateral (*LL*: LLPf, LLFr, LLTe, LDPa, LLOc), left medial (*LM*: LMPf, LDFr, LMFr, LMCe, LMOc), right lateral (*RL*: RLPf, RLFr, RLTe, RDPa, RLOc) and right medial (*RM*: RMPf, RDFr, RMFr, RMCe, RMOc) electrode sets that included congruence (match vs. mismatch) and anteriority (5 levels). Given are the F-values; we report main effects of congruence (*C*) and interactions of congruence with hemisphere (*H*), laterality (*L*), and anteriority (*A*); main effects of factors hemisphere, laterality, and anteriority are omitted for the sake of brevity; degrees of freedom *df*(1,23) expect for CA, CHA, CLA, CHLA, *df*(4,20).
*$p < .05$; **$p < .01$; ***$p < .001$.

Table 2. *Statistical Results Mean Amplitude of ERPs at the Second Noun*

Sentence position	Time window	Factors	Overall ANOVA	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
2nd noun (patient)	0–100	C	—	—	—	—	—
		CH					
		CL					
		CA	3.85*	5.38*	—	—	—
		CHL	—				
		CHA	—				
		CLA	—				
		CHLA	—				
	100–300		—				
	300–500	C	—				
		CH	5.05*				
		CL	—				
		CA	—				
		CHL	—				
		CHA	—				
		CLA	—				
		CHLA	—				
	400–600	C	—				
		CH	5.18*				
		CL	—				
		CA	—				
		CHL	—				
		CHA	—				
		CLA	—				
		CHLA	—				

Note: See Table 1 for legend.
*$p < .05$.

higher the mean accuracy scores, the larger the N400 congruence effect (Table 3 and Figure 6b).

For the negativity at the second noun (300–500 ms), there was a trend showing that the larger the frontolateral N400 effect, the larger the response latency congruence effect. This relation was reliable for a slightly later (400–600 ms) time window, over the right lateral slice only (see Table 3), and not at individual electrode sites ($-0.37 < r < 0.11$, see Figure 4). Correlations for ERPs to the second noun with either verb ERPs, or with the other behavioral measures, were not reliable.
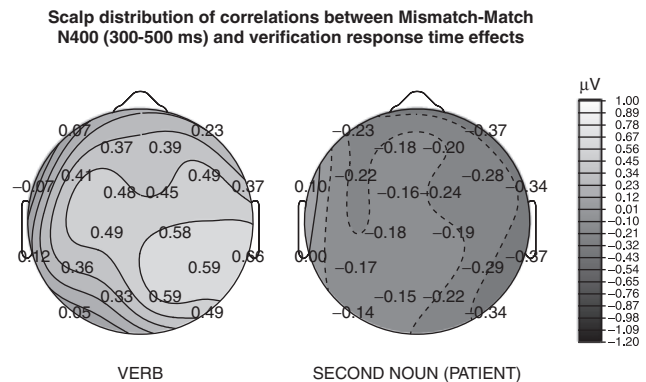
In sum, participants differ systematically as to when they process the congruence: those who exhibit a large response time verification effect (e.g., 3, 10, 17, 19, 15, 22, and 9, see Figure 5) exhibit a small, or even reverse, congruence N400 at the verb (positive numbers on the x-axis reflect participants for whom incongruous verbs were more positive than congruous verbs). In contrast, participants that exhibit a small response time verification effect (e.g., 20, 6, 23, 8, 11, 14, 21) display a medium to large congruence N400 (i.e., a negative number on the x-axis; mismatching trials were more negative than matching trials) at the verb. Participants with large "early" congruence effects also tended to have higher accuracy (Figure 6b). Post-verbal ERP congruence effects were positively correlated with the response congruence effects: Participants that exhibit a large congruence
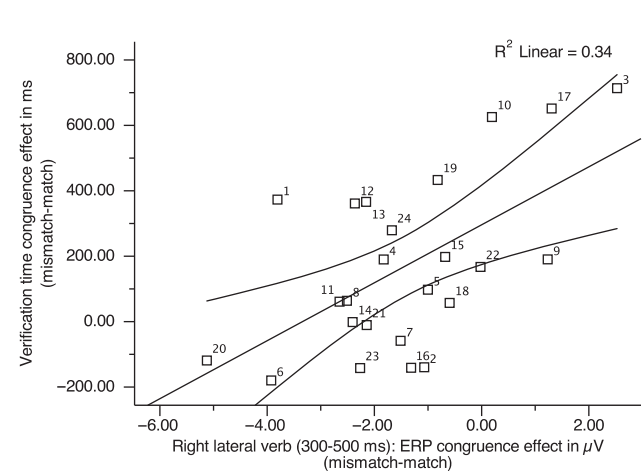

Figure 4. The *r*-values for correlations end-of-sentence response latency difference scores (mismatch minus match) and mean ERP amplitude difference scores at 20 electrodes (mismatch minus match) for the verb and the second noun in the 300–500 time windows. Lighter shades indicate larger positive correlations (i.e., a negative relationship) between verb and verification time congruence effects; darker shades depict negative correlations between second noun and verification time congruence effects. The *r*-values between individual sites were estimated by spherical spline interpolation.

Figure 5. Scatter plots of correlation coefficients between response latency difference scores and ERPs summarized for each participant: right lateral slice N400 verb.

Table 3. *Correlation Analyses for the Verb and Second Noun*

	ERP congruence effect			Behavioral measures	
	Left medial sites	Right lateral sites	Right medial sites	RT congruence effect	VWM scores
Verb (300–500 ms)					
Verification mean accuracy	—	—	− 0.47*	− 0.45*	0.38#
RT congruence effect	0.49*	0.58**	0.57**		− 0.47*
VWM scores	—	—	—	− 0.47*	
ECFT scores	—	—	—	—	—
Second noun (400–600 ms)					
Verification mean accuracy	—	—	—		
RT congruence effect	—	− 0.49*	—		
WVM scores	—	—	—		
ECFT scores	—	—	—		

Note: #$p < .1$.
*$p < .05$; **$p < .01$.

N400 at the second noun exhibit a large response time congruence effect; participants with a small congruence effect at the second noun exhibit a small response time congruence effect.

The time course and accuracy with which a participant processes congruence further appears to be a function of his/her verbal working memory scores: the higher a participant's verbal working memory score, the smaller his/her response latency

Figure 6. Scatter plots of correlation coefficients (a) between response latency difference scores and verbal working memory scores; and (b) between (right medial) verb ERP congruence effects and percentage of mean response accuracy for incongruous and congruous trials.

congruence effect and vice versa (see Table 3 and Figure 6a); and individuals with higher mean accuracy scores had a small, if any, response time congruence effect while those with lower mean accuracy scores had a large one. A participant's verbal working memory score also correlated with his/her mean accuracy (the higher the VWM score, the higher mean accuracy, Table 3). Visual-spatial (ECFT) test scores, in contrast, were not reliably correlated with any ERP or verification time congruence effects.
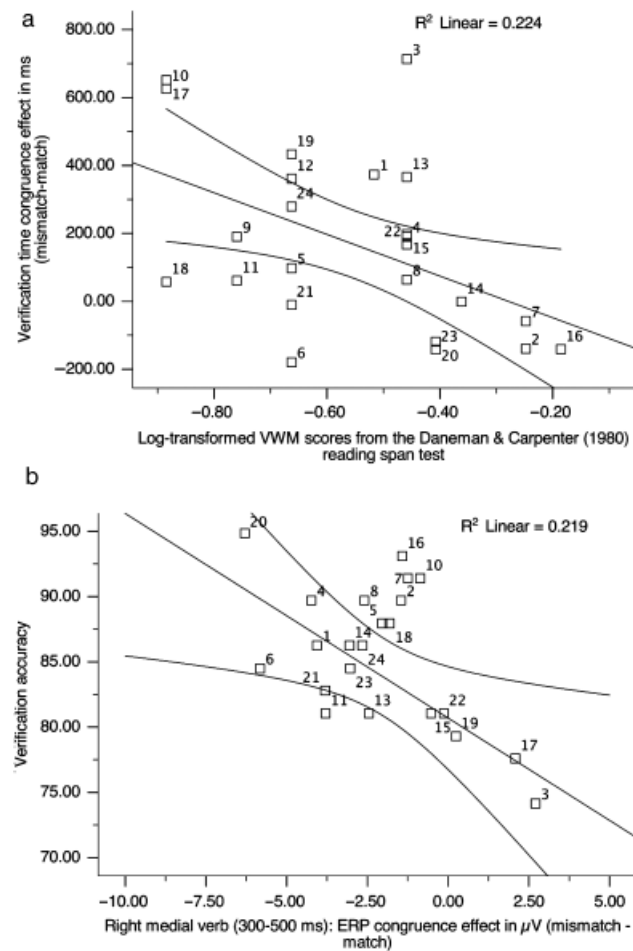
Standard Simple and Multiple Regression Analyses

Next, we computed simple regression analyses (e.g., right lateral verb N400s), and also multiple regression analyses with ERPs to the verb and post-verbal regions (e.g., right lateral verb N400s, and right lateral ERP effects at the second noun) to examine how well these variables predict the response time congruence effect. We systematically explored a number of models (different combinations of mean amplitude ERP difference scores at the verb, object noun, and verbal working memory scores) and computed partial Fs to find the best-fitting model. The best-fitting model was one that combined right (but not left) lateral verb N400s and second noun N400s with VWM scores as predictors (right lateral: $(F1(3,20) = 11.11$, $p < .001$, $r = 0.79$, adjusted $R^2 = 0.57$, coefficient verb: $t = 2.82$, $p < .05$; coefficient second noun: $t = − 3.04$, $p < .05$; coefficient VWM: $t = − 2.66$, $p < .05$). This model accounted for substantially more variance than combinations of right lateral verb and second noun models (adjusted $R^2 = 0.44$, partial $F(1,21) = 4.64$, $p < .05$), and the latter accounted again for more variance than simple (e.g., right lateral) verb models ($r^2 = 0.34$, adjusted $R^2 = 0.31$, partial $F(1,21) = 4.88$, $p < .05$).

Follow-up ANOVAs: VWM

To more fully examine the relationship between ERP and response latency congruence effects, we analyzed data based on a median split of participants' VWM scores (informed by visual inspection).[1] Specifically, we performed repeated measures

---

[1]VWM score for one participant—while just below the median of − 0.4876—clustered (by eyeball and k-means clustering) with the above rather than below VWM median scores. Since k-means clustering results depend on the ordering of the scores, we performed clustering on 10 randomizations of participants' VWM scores. Membership of participants in the high versus low VWM score cluster was highly consistent across multiple runs on randomized scores. The 'low VWM' cluster always had 11 participants and a final center at around − 0.79; the 'high VWM' cluster contained scores from the remaining 13 participants with the final center around − 0.40. The difference between the two groupings
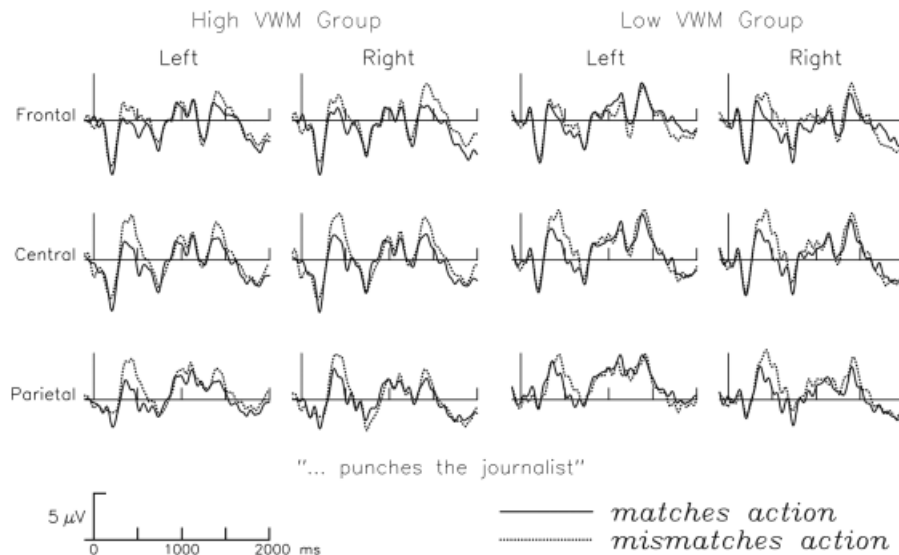
Figure 7. Grand average ERPs (mean amplitude) at the verb and post-verbal regions for participants with high and low verbal working memory scores.

ANOVAs on the ERP and verification congruence scores separately for each VWM group. For the ANOVAs, we ensured that the data in the two groups were approximately normally distributed by inspecting skew and kurtosis values both separately for each group and across the two groups.

ANOVAS for low and high VWM groups. Analyses revealed—as expected—a reliable verification time congruence effect only for the low VWM group ($FI(1,10) = 8.58$, $p < .02$, $\eta^2 = 0.46$; mean time congruous 1104 ms: mean time incongruous: 1337 ms; high VWM group: $p > .1$, $\eta^2 = 0.17$; mean time congruous: 1105 ms mean time incongruous: 1218 ms). Across conditions, the high VWM group's average response time was faster (1162 ms) than the low VWM group's average response time (1221 ms).

Visual inspection of grand average difference waves (mismatch minus match) for the two VWM subgroups suggested a larger P2 difference for the high than for the low VWM group. The N400 congruence effect at the verb was only slightly larger for the high VWM group but seemed to arise earlier, and the more negative-going waveforms for the higher VWM participants increased during the post-verbal determiner and second noun especially over frontal and lateral sites (see Figure 7).

In the ERP analyses, a reliable main effect of congruence between 180–220 ms, (the P2 region), for the high ($FI(1,12) = 8.5$, $p < .02$, $\eta^2 = 0.41$) but not for the low ($F < 1$) VWM group confirmed the visual impression. Both groups showed reliable verb N400 congruence effects between 300–500 ms (for low VWM, $FI(1,10) = 5.77$, $p < .05$, $\eta^2 = 0.37$; for high VWM, $FI(1,12) = 17.28$, $p < .01$, $\eta^2 = 0.59$). However, only the high VWM group showed a reliable main effect of congruence in the early half of the verb N400 (300–400 ms: $FI(1,12) = 8.72$, $p < .02$, $\eta^2 = 0.42$). For the second noun (300–500 ms), analyses revealed a reliable congruence by hemisphere interaction for the high VWM group only ($FI(1,12) = 5.00$, $p < .05$, $\eta^2 = 0.29$; low VWM group, $F < 1$).

(clustering vs. median split) concerned only the one participant that was just below the median VWM but grouped more naturally with the above VWM group.

Discussion

In this study, we assessed the suitability of the verification task—insofar as it reveals visual context influences on sentence comprehension—as a paradigm for investigating situated language comprehension. We recorded ERPs along with post-sentence RTs during a picture-sentence verification task in which participants judged whether a simple subject-verb-object sentence, presented visually one word at a time, matched the preceding depicted action event involving two characters. We examined (a) whether the effect of the visual scene on sentence processing is just immediate and local (i.e., limited to the verb), and thus explicable in terms of simple priming, or manifest at multiple time points during sentence processing; (b) the relationships between verification time and ERP congruence effects (via correlations and linear regressions); and (c) the relationship of each of these congruence effects with VWM scores. From these, we determine the nature of the relationship between visual context influences on online comprehension and end-of-sentence verification times, in order to infer the time course of verification-related processes.

Overall, the observed data pattern provides strong evidence against a simple priming account. First, we observed multiple reliable ERP congruence effects over the course of the sentence: At the earliest possible point in the sentence, namely the verb (300–500 ms), N400 amplitudes over centro-parietal scalp were larger for verb-action mismatches than matches. The centro-parietal N400 effect was indistinguishable from that of lexico-semantic anomalies or low cloze probability words in sentences read for comprehension (e.g., Kutas, 1993; see also Kutas, Van Petten, & Kluender, 2006; Otten & van Berkum, 2007; Van Berkum et al., 1999). We thus conclude that the verb N400 in the present study reflects visual context effects on language comprehension. Crucially, we also observed picture-sentence mismatch effects post-verbally, in the form of a negativity in response to the object noun (between 300–500 and 400–600 ms). Compared to the N400 at the verb, this negativity had a more anterior scalp distribution. While the functional significance of this post-verbal negativity is unclear (see Wassenaar & Hagoort, 2007 for comparison), the topographic difference between the verb and object noun ERP effects implicates at least two distinct congruence

effects during the processing of these short sentences. It also provides some evidence against any simple action-verb priming account, which by itself provides no principled explanation for multiple, distinct ERP effects. Visual context effects also emerged, as expected, in post-sentence RTs, when participants gave their verification responses, with longer latencies for picture-sentence mismatches than matches.

These multiple congruence effects—the verb N400 (reflecting visual context effects on comprehension), the object noun negativity (indexing further aspects of congruence processing), and the post-sentence RTs (widely presumed to index verification processes)—could, in principle, reflect the same (e.g., priming) processes; if so, we would expect these measures to co-vary directly. The pattern of correlations, however, is more complex than any single mechanism (e.g., priming) could readily accommodate: (1) there is an inverse correlation between the verb N400 congruence effect and the verification RT congruence effect: the larger the N400 congruence effect, the smaller the RT congruence effect; (2) the object noun ERP congruence effect, on the other hand, does correlate directly with the RT congruence effect: the larger the object noun congruence effect, the larger the RT congruence effect; and (3) moreover, the verb N400 and object noun congruence effects do not reliably correlate with each other, as might be expected if they were both consequences of the same cause.

Multiple regression analyses corroborate the case against a simple priming account. The best predictor of verification time effects in a linear regression proved to be a combination of the verb N400 congruence effect plus the second noun congruence ERP effect (from right but not left hemisphere sites) plus verbal (and not visual) working memory scores. These three measures together account for 57% of the variance in verification RT effects. This is significantly greater than the variance accounted for either by the simple verb N400 effect ($\sim 31\%$) or even by the significantly greater variance accounted for by the combination of the verb N400 congruence effect plus the second noun congruence effect ($\sim 44\%$). Verbal working memory scores on their own also were a reliable predictor of the RT effects, but not of the N400 congruence effects. Nonetheless, separate analyses confirmed earlier congruence effects in the high span comprehenders (P2 and early N400 region) relative to the low span comprehenders (late N400 region and in the RTs). In the absence of any correlations with visual-spatial scores from the Extended Complex Figure Test, there is thus no evidence that the time course differences are related to participants' differential abilities in recognizing the depicted actions. We speculate that the high VWM group may have used the visual scene more rapidly as a source of contextual information. This account is compatible with the proposal that individuals with high verbal working memory may be more motivated and/or attentive (see Chwilla, Brown, & Hagoort, 1995).

To reiterate, these results taken all together cannot be accounted for by any simple priming account. Furthermore, they also run counter to any view that relegates all verification-related processes to some time separate from, and after, comprehension (Tanenhaus et al., 1976). The exact timing of verification-related processes that contribute to the post-sentence verification times does seem to vary across individuals as a function of (at least) their verbal working memories. These individual differences not withstanding, the data are clear in demonstrating that verification-related processes, like comprehension processes, are distributed in time and incremental in nature. Indeed, the first ERP indicant of a picture-sentence mismatch not only occurs at the earliest moment it could (at the verb) in a component that is known to co-vary with comprehension (i.e., posterior N400) but is also predictive of post-sentence verification times. A qualitatively different contribution to verification RTs is evidenced in the ERP congruity effects at the second noun.

Our findings thus support the methodological criticism that end-of-sentence verification times alone underdetermine models of the time course of congruence processing in such verification tasks (see Kounios & Holcomb, 1992; Tanenhaus et al., 1976). It is only the combination of continuous measures (e.g., eye tracking or ERPs) and RTs that permits us not only to validate the paradigm but to better delineate the time course of visual context effects. The multiple systematic relationships between a participant's incremental ERP congruence effects and their verification response time pattern at sentence end indicate that overt picture-sentence verification decisions post sentence are some function of how picture-sentence congruence is processed during incremental sentence comprehension, at least when people are encouraged to understand both the pictures and sentences. The variation of these effects with verbal working memory scores shows further that individual differences play a systematic role in the cognitive processes at work here, as elsewhere in sentence comprehension (e.g., Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Just & Carpenter, 1992; Waters & Caplan, 2001).

A better understanding of comprehension-verification relationships is important for theories of (situated) language comprehension, since verification (overt responses on correspondence between what is said and how things are) appears to be part and parcel of routine language communication, and is thus a potentially pervasive mechanism. Positive verification is evident in expressions of agreement ("So I heard," "No doubt") while failures to verify may be inferred from corrections and expressions of disbelief, uncertainty, requests for clarification, and the like (e.g., "Well no, actually what happened was . . .", "Are you sure?").

In sum, the present data re-establish the validity of the picture-sentence verification paradigm for the study of online language comprehension—provided post sentence RTs are complemented by continuous measures—thereby paving the way for investigations of comprehenders' reliance on representations of linguistic and non-linguistic visual context even when they do not perfectly match those of the current linguistic input.

REFERENCES

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The blank screen paradigm. *Cognition, 93*, B79–B87.

Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition, 14*, 201–212.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic model of verification. *Psychological Review, 82*, 45–76.

Carpenter, P. A., & Just, M. A. (1976). Models of sentence verification and linguistic comprehension. *Psychological Review, 83*, 318–322.

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

Chwilla, D. J., Brown, C., & Hagoort, P. (1995). The N400 as a function of the level of processing. *Psychophysiology, 32*, 274–285.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3, 472–517.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12, 769–786.

Dale, A. M. (1994). Source localization and spatial discriminant analysis of event-related potentials: Linear approaches. Unpublished doctoral dissertation, University of California, San Diego.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 3, 422–433.

D'Arcy, R. C. N., & Connolly, J. F. (1999). An event-related brain potential study of receptive speech comprehension using a modified Token Test. *Neuropsychologia*, 37, 1477–1489.

Fasteneau, P. S. (2003). *Extended complex figure test*. Los Angeles, CA: Western Psychological Services.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41, 469–495.

Federmeier, K. D., & Kutas, M. (2002). Picture the difference: Electrophysiological investigations of picture processing in the cerebral hemispheres. *Neuropsychologia*, 40, 730–747.

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. J. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, 20, 400–409.

Goolkasian, P. (1996). Picture–word differences in a sentence verification task. *Memory and Cognition*, 24, 584–594.

Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning & Verbal Behavior*, 4, 107–111.

Griffin, D., Murray, S., & Gonzalez, R. (1999). Difference score correlations in relationship research: A conceptual primer. *Personal Relationships*, 6, 505–518.

Holcomb, P. J., & Neville, H. J. (1991). Natural speech processing: An analysis using event-related brain potentials. *Psychobiology*, 19, 286–300.

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics by using generalized linear models*. Thousand Oaks, CA: Sage.

Johns, G. (1981). Difference score measures of organizational behavior variables: A critique. *Organizational Behavior and Human Performance*, 27, 443–463.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122–149.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role assignment: Evidence from eye movements in depicted events. *Cognition*, 95, 95–127.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science*, 30, 481–529.

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

Knoeferle, P., Habets, B., Crocker, M. W., & Münte, T. F. (2008). Visual scenes trigger immediate syntactic reanalysis. *Cerebral Cortex*, 18, 789–795.

Kounios, J., & Holcomb, P. J. (1992). Structure and process in semantic memory: Evidence from event-related brain potentials and reaction times. *Journal of Experimental Psychology: General*, 121, 459–479.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single word versus sentence context effects. *Language and Cognitive Processes*, 8, 533–572.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207, 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307, 161–163.

Kutas, M., & Van Petten, C. (1988). Event-related brain potential studies of language. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (*Vol. 3*, pp. 139–187). Greenwich, Connecticut: JAI Press.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II. In M. A. Gernsbacher & M. Traxler (Eds.), *Handbook of psycholinguistics* (2nd edition, pp. 659–724). New York: Elsevier Press.

Mayberry, M., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*, 33, 449–496.

McCallum, W., Farmer, S. F., & Pocock, P. V. (1984). The effects of physical and semantic incongruities on auditory event-related potentials. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 5, 477–488.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 227–234.

Otten, M., & Van Berkum, J. J. A. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario fit on the discourse-dependent N400 effect. *Brain Research*, 1153, 166–177.

Reichle, E. D., Carpenter, P. A., & Just, M. A. (2000). The neural basis of strategy and skill in sentence-picture verification. *Cognitive Psychology*, 40, 261–295.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109–148.

Singer, M. (2006). Verification of text ideas during reading. *Journal of Memory and Language*, 54, 574–591.

Spivey, M., Tanenhaus, M. K., Eberhard, K., & Sedivy, J. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45, 447–481.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.

Szűcs, D., Soltész, F., Czigler, I., & Csépe, V. (2007). Electroencephalography effects to semantic and non-semantic mismatch in properties of visually presented single-characters: The N2b and the N400. *Neuroscience Letters*, 412, 18–23.

Tabachnik, B., & Fidell, L. S. (2007). *Using multivariate statistics*. Pearson Allyn & Bacon: Boston.

Tanenhaus, M. K., Carroll, J. M., & Bever, T. G. (1976). Sentence-picture verification models as theories of sentence comprehension: A critique of Carpenter and Just. *Psychological Review*, 83, 310–317.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology*, 56, 165–182.

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, 11, 657–671.

Vissers, C., Kolk, H., Van de Meerendonk, N., & Chwilla, D. (2008). Monitoring in language perception: Evidence from ERPs in a picture-sentence matching task. *Neuropsychologia*, 46, 967–982.

Wassenaar, M., & Hagoort, P. (2007). Thematic role assignment in patients with Broca's aphasia: Sentence-picture matching electrified. *Neuropsychologia*, 45, 716–740.

Waters, G. S., & Caplan, D. (2001). Age, working memory, and online syntactic processing in sentence comprehension. *Psychology and Aging*, 16, 128–144.