Different mechanisms for role relations versus verb-action congruence effects: Evidence from ERPs in picture-sentence verification

Pia Knoeferle
(Cognitive Interaction Technology Excellence Cluster Bielefeld University)
Thomas P. Urbach
(Department of Cognitive Science University of California San Diego)
Marta Kutas
(Department of Cognitive Science University of California San Diego)

Correspondence regarding this article should be addressed to: Dr. Pia Knoeferle,
Inspiration 1, FBIIS, Room 2.036
Bielefeld University
D-33615 Bielefeld
Germany
Email: knoeferl@cit-ec.uni-bielefeld.de
Fax: +49 521 106-6560

Abstract

Extant accounts of visually situated language processing do make general predictions about visual context effects on incremental sentence comprehension; these, however, are not sufficiently detailed to accommodate potentially different visual context effects (such as a scene-sentence mismatch based on actions versus thematic role relations, e.g., Altmann & Kamide, 2007; Knoeferle & Crocker, 2007; Taylor & Zwaan, 2008; Zwaan & Radvansky, 1998). To provide additional data for theory testing and development, we collected event-related brain potentials (ERPs) as participants read a subject-verb-object sentence (500 ms SOA in Experiment 1 and 300 ms SOA in Experiment 2), and post-sentence verification times indicating whether or not the verb and/or the thematic role relations matched a preceding picture (depicting two participants engaged in an action). Though incrementally processed, these two types of mismatch yielded different ERP effects. Role-relation mismatch effects emerged at the subject noun as anterior negativities to the mismatching noun, preceding action mismatch effects manifest as centro-parietal N400s greater to the mismatching verb, regardless of SOAs. These two types of mismatch manipulations also yielded different effects post-verbally, correlated differently with a participant's mean accuracy, verbal working memory and visual-spatial scores, and differed in their interactions with SOA. Taken together these results clearly implicate more than a single mismatch mechanism for extant accounts of picture-sentence processing to accommodate.

Keywords: situated language processing accounts; sentence-picture verification; visual context effects; event-related brain potentials

---

Introduction

Language processing is central to a diverse range of communicative tasks including reading books, exchanging ideas, and watching the news, among many others. It also plays an important role in tasks in which communication is not the primary goal such as navigating in space, buying a ticket at a vending machine, or acquiring new motor skills. Indeed, much language processing takes place in a rich non-linguistic context. Such 'situated' language comprehension has been investigated in a variety of tasks using a variety of dependent measures including response times, eye movements, and event-related brain potentials (ERPs) - studies from which a reliable set of findings has emerged.

Perhaps most notably it has become increasingly clear that language is robustly mapped onto visual context. Incongruence (vs. congruence), for instance, affects how rapidly people verify a written sentence against a picture - faster for matching than mismatching stimuli (e.g., Clark & Chase, 1972; Gough, 1965). Moreover, it does so whether the task is verification (in response times) or sentence reading (in fixation times precisely at the word that mismatches aspects of the visual context, Knoeferle & Crocker, 2005).

There is also a general consensus that different aspects of a situational context - such as space, time, intentionality, causation, objects, protagonist - contribute to the construction of mental representations/models (see Zwaan & Radvansky, 1998, for a review). Modifications to each of these aspects can engender longer response times to probes and/or total sentence reading times, when there is a change in time or place in a narrative versus when there is not.

Another seminal finding concerns the time course of language-vision integration and the role of the visual context in language processing. The pattern of eye movements to objects as participants listen to related sentences in the 'visual world paradigm' has shown that a referential visual context can help resolve linguistic ambiguity within a few hundred milliseconds (e.g., Tanenhaus et al., 1995). This paradigm also has been used to argue that people anticipate objects when the linguistic context is sufficiently constraining (Altmann & Kamide, 1999; Kamide et al., 2003; Sedivy et al., 1999).

Similarly, anticipation seems to occur when spoken sentences are ambiguous but action events impose constraints on visual attention (Knoeferle et al., 2005). In short, language processing is temporally coordinated with visual attention to objects and events, presumably enabling rapid visual context effects on comprehension [1].

Last but not least, recent studies suggest that these visual context effects involve functionally distinct comprehension processes. Knoeferle et al. (2011), for instance, examined verb-action relationships as participants read a subject-verb-object sentence and verified whether or not the verb matched an immediately preceding depicted action. Two qualitatively distinct ERP effects emerged (one at the verb, the other at the post-verbal object noun) implicating functionally distinct processes in understanding even a 'single' (verb-action) mismatch. Mismatches between the role relations expressed in a sentence and depicted in a drawing elicited at least partially different ERP effects from those to verb-action mismatches, thereby corroborating the hypothesis that there may be functionally distinct mechanisms in mapping language to the visual context (Wassenaar & Hagoort, 2007, although see Vissers et al. (2008)).

Accounting for visual context effects in language comprehension

Results such as these from situated language research have inspired a host of models that differ in their coverage (of a specific task or language comprehension more generally), their natures (frameworks vs. processing accounts), and their representational assumptions (modular or not). Among the task-specific models, the 'Constituent Comparison' model accommodates picture-sentence verification (Carpenter & Just, 1975), whereas the 'Moni-

---

[1] By 'visual context effects' we mean the influence of scene-derived representations on language comprehension processes.

toring Theory' accommodates error monitoring (see, e.g., Kolk et al., 2003; Van de Meerendonk et al., 2009). These two models, however, provide limited coverage of comprehension more broadly and have proven inadequate as they predict no incremental effects (Constituent Comparison Model) or the same response to any type of incongruence (Monitoring Theory).

With regard to their nature, some accounts (e.g., situation models) are best characterised as frameworks for the construction of mental models in language and memory tasks. One situation model - the event-indexing model, for instance - specifies that a newly incoming cue (e.g., a new protagonist) leads to an update of the relevant index (e.g., the protagonist index). This model is underspecified as to precisely when such updates occur and how they might affect specific comprehension processes. Other accounts, by contrast, specifically designed to accommodate the processes implicated in real-time situated language comprehension (e.g., Altmann & Kamide, 2009; Knoeferle & Crocker, 2006, 2007), all assume rapid influences of non-linguistic representations on language processing but differ in their representational and mechanistic assumptions.

Altmann & Kamide (2009), for example, postulate a single representational format for different aspects of language and the visual context, in line with recent accounts of embodied cognition (Barsalou, 1999, see also Glenberg & Robertson (1999)). By contrast, the Coordinated Interplay Account (CIA, Knoeferle & Crocker, 2007) assumes distinct language and scene-derived representations. With regard to mechanisms, some of these accounts assume that corresponding elements of a sentence and of the visual context in the focus of attention are co-indexed, thereby establishing reference (Glenberg & Robertson, 1999; Knoeferle & Crocker, 2006, 2007). Others postulate attention-mediated representational overlap and competition among representations (Altmann & Kamide, 2009).

These situation models cover a broad range of situational dimensions and associated mental representations, but are underspecified regarding the real-time coordination of language processing and (visual) attention. The various real-time processing accounts, by contrast, can at least in broad stroke, accommodate the rapid coordination of language processing, visual attention, and visual context effects but are underspecified with regard to how different aspects of the situation model feed into distinct comprehension processes (but see Crocker et al., 2010). In sum, there is no principled account of how visual context affects functionally distinct processes during situated comprehension. Indeed, we have limited knowledge of the relative time courses or types of processes underlying the different visual context effects during language comprehension, although these are clearly key to any account of how language is interpreted against a current visual background (i.e., situated language comprehension). In the present studies we aim to help fill this theoretical gap by collecting ERPs to distinctly different sorts of picture-sentence mismatches in a verification task.

Verb-action versus thematic role relations mismatches: ERPs and RTs

Specifically, we conducted two picture-sentence verification studies, each with two different types of violation within individuals. We used a known (verb-action) mismatch that elicits an N400 to the mismatching verb and negativity to the patient noun and a role relations mismatch. Given a picture of a gymnast punching a journalist, a sentence such as *The gymnast punches* *the journalist* constitutes a complete match; a sentence about the gymnast applauding the journalist includes a verb-action mismatch; a sentence about the journalist punching the gymnast includes a role-relation mismatch; and a sentence stating the journalist applauds the gymnast includes both a role-relation mismatch (wrong agent and patient) and an action mismatch (wrong action).

We recorded ERPs as participants inspected one of these types of pictures and shortly thereafter read an NP1-Verb-NP2 sentence, after which we collected their end-of-sentence verification response. To aid in our interpretation of the (mis)match effects, we also collected participants' scores in the reading span test (Daneman & Carpenter, 1980) and a motor-independent version of the extended complex figure test (Fasteneau, 2003). We compared the ERPs to the two violation types in morphology, timing, and scalp topography. We also examined their relationships to end of sentence responses and to other behavioural variables (e.g., verbal and visual-spatial working memory). We plan to use the extent to which action and depicted role relation (mismatch) effects on language comprehension are similar in these respects to determine whether or not the effects are best accounted for by a single functional cognitive / neural mechanism or more.

Predictions of a single cognitive / neural mechanism. If a single mechanism is engaged by any mismatch between the pictorial representation and the ensuing verbal description, then any and all mismatches should elicit the same ERP response, though they might differ in timing. Participants may assign roles to the depicted event participants (e.g., a patient role to the gymnast) and compare these to sentential role relations as they read a sentence (e.g., *The gymnast applauds...*). Depending on when they assign the thematic (agent) role to the first noun phrase, this may occur as soon as the first noun, or perhaps not until the verb. If both the role relations and verb action mismatch effects appear at the verb, they may be indexed by larger negative mean amplitude ERPs compared with matches (N400) as reported for active sentences (Knoeferle et al., 2011; Wassenaar & Hagoort, 2007).

Moreover, if all mismatches engage the same cognitive / neural mechanism, we would expect them to co-vary similarly with behavioral measures. We have reported reliable correlations between N400 congruence effects at the verb and end-of-sentence congruence response times in young adults (Knoeferle et al., 2011). Participants with a small N400 congruence effect at the verb tended to exhibit a large response time congruence effect at sentence end, and vice versa. In addition, participants with lower verbal working memory tended to have larger response time congruence effects, suggesting that the time course of congruence processing might vary with verbal working memory. With the present study we can see whether these findings will replicate and/or generalize, and the extent to which role relation and action mismatches behave similarly. Under a single mechanism view, we should see mismatch effects in the response times to action and role-relation mismatches alike.

Predictions: More than a single cognitive / neural mechanism?. Alternatively, if more than a single mechanism subserves various picture-sentence mismatches, then we aim to deduce their natures and relative timing from the ERP and RT data. N400 effects, for example, are usually taken to reflect semantic processing and contextual relations (see Kutas & Federmeier, 2011, for a review). Some of the reported between-experiment variation in congruence processing in the literature may reflect the sensitivity of the N400 and/or ERPs more generally to different types of mismatches. Extant studies, however, also differ in other ways: spoken comprehension in healthy older adults (Wassenaar & Hagoort, 2007) versus sentence reading in younger adults (Knoeferle et al., 2011).

If these reported results (Knoeferle et al., 2011; Wassenaar & Hagoort, 2007) replicate within subjects in the same experiment, we would expect to see larger N400s at the verb and post-verbal noun for both verb-action and sentence role relations mismatches relative to matches, a post-N400 positivity for the role relations mismatches only, and end-of-sentence response time mismatch effects for the verb-action mismatches only. Moreover, to the extent that the N400 mismatch effects and the relative positivity reflect functionally distinct neural processes, we would expect them to correlate differently with end-of-sentence RTs and the behavioral scores.

If we see N400 amplitude modulations at the verb for both kinds of mismatches, there are several possible outcomes. If these two kinds of mismatches are processed by separate stages (as in any strictly serial account), both of which contribute to verb processing, then the N400 amplitude would reflect additivity (see, e.g., Hagoort, 2003; Kutas & Hillyard, 1980; Sternberg, 1969, for the methodology and its application to ERP data): Double mismatches would yield the largest N400s and the longest end-of-sentence response times, no mismatches the smallest N400 amplitudes and the shortest RTs, and single matches of either kind intermediate N400s. Alternatively, if these two types of mismatches engage interacting processes, we would expect to see non-additivity in the ERPs and RTs. The CIA (like other accounts) is underspecified as to whether or not these two mismatches might interact.

Alternatively, verb-action and role relations mismatches may not emerge at the same word (the verb). On a fully incremental account, participants could assign a patient role to the gymnast upon seeing the gymnast as the patient in an event depiction, an agent role to it upon reading the noun phrase *the gymnast*, in sentence-initial position (see, e.g., Bever, 1970), and thus immediately experience a mismatch. On this possibility, the ERPs might index role mismatch at the first noun, before a verb-action mismatch; on the assumption that a mismatch earlier in the sentence enables earlier preparation and thus faster response execution, response times to role relations would be faster than those to action mismatches. These role mismatch effects might manifest as an N2b, as observed for adjective-color mismatches (D'Arcy & Connolly, 1999) and role relations mismatches in irreversible active sentences (Wassenaar & Hagoort, 2007), or as a N400-like relative negativity to the first noun.

Another, albeit less likely, alternative given the incremental and/or predictive nature of language comprehension (e.g., Elman, 1990; Federmeier, 2007; Hale, 2003; Kamide et al., 2003; Levy, 2008; Pickering & Garrod, 2007) is that the depicted and sentential role relations are compared only *after* people have accessed the verb's lexical entry (e.g., Carlson & Tanenhaus, 1989; MacDonald et al., 1994). If so, then we might expect to see later ERP effects and perhaps slower verification times for role than verb-action mismatches.

In sum, we believe that the relative time course of the congruence effects to these two picture-sentence mismatch manipulations, their topographies and their relationships with end-of-sentence verification latencies and neuropsychological test scores, will provide additional constraints on (single or more) mechanisms found in accounts of visually situated language comprehension.

Experiments 1 and 2

Methods

Participants

Thirty-two students of UCSD took part in Experiment 1 (16 females, 16 males; aged 18-29, mean age: 20.84); a different set of thirty-two participated in Experiment 2 (16 females, 16 males; aged 18-23, mean age = 19.94). All participants were native English speakers, right-handed (Edinburgh Handedness Inventory), and had normal or corrected-to-normal vision. All gave informed consent; the UCSD IRB approved the experiment protocol.

Materials, design, and procedure

Materials for both experiments were derived from Knoeferle et al. (2011) by creating two new pictures and sentences for each item. The design by Knoeferle et al. (2011) had 1 within-subjects factor (*action congruence* with the levels congruent, Picture 1a vs. incongruent, Picture 1b, see Table 1). To this we added Pictures 1c and 1d, resulting in a 2 x 2 within-subject design with the factors *role-relation congruence* (congruent, Picture 1a/b vs. incongruent, Picture 1c/d) and *action congruence* (Pictures 1a/c vs. 1b/d, Table 1).

The sentence, *The gymnast punches the journalist*, in Table 1 is congruent with respect to both the action and role dimensions for Picture 1a,

Table 1: about here

(full match); it is incongruent with respect to the action but congruent with respect to the role-relations dimension for Picture 1b (action mismatch); it is congruent with respect to the action but incongruent with respect to the role relations dimension with Picture 1c (role mismatch); and it is incongruent with respect to both these dimensions for Picture 1d (combined mismatch). In Knoeferle et al. (2011) 21 of the 80 items 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 was changed for the present experiments such that only simple noun phrases were used.

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) occurred once in a congruent (match) and once in an incongruent (mismatch) 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) were also counterbalanced.

There were 80 item sets which, combined with the conditions and the counterbalancing (counterbalancing measures (1) and (3)), yielded 16 experimental lists. Each list contained one occurrence of an item, and an equal number of left-to-right and right-to-left action depictions. Each list also contained 160 filler items, of which half were mismatches. These filler sentences had different syntactic structures including negation, clause-level and noun phrase coordination, as well as locally ambiguous reduced relative clause constructions in which the first noun phrase was the patient of the reduced relative clause. The fillers also ensured that a sentence initial noun phrase was not always a felicitous agent. For some fillers the sentence started with a noun phrase but the picture was fully unrelated; and for other fillers, the first-mentioned noun phrase mismatched the picture referentially.

Procedure

Participants inspected the picture on a CRT monitor for a minimum of 3000 ms terminated via a right thumb button press. Next, a fixation dot appeared for a random duration between 500 and 1000 ms, followed by the sentence, one word at a time. Word onset asynchrony was 500 ms in Experiment 1 and 300 ms in Experiment 2; word presentation duration was 200 ms in both. Participants were instructed to examine the picture and then to read and understand the sentence in the context of the preceding picture. Participants indicated via a button press as quickly and accurately as possible after each sentence whether it matched the preceding picture or not. After that button press, there was a delay interval randomly varying between 500 and 1000 ms prior to the next trial.

Recording and analyses

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 1) 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 analyses (unless otherwise stated) were conducted relative to a 200-ms pre-stimulus baseline. All trials were scanned offline for artefacts, and contaminated trials were excluded from further analyses. Blinks were corrected with an adaptive spatial filter (Dale, 1994) for 20 of the participants in Experiment 1, and 12 participants's data in Experiment 2. After blink correction, we verified that less than 27% of the data for a given participant per condition at a given word region were rejected. However, artefact rejection rates per condition after blink correction were higher than 27 % for 2 participants at the first noun, 3 at the verb, and 11 participants at the second noun in Experiment 1, and for 2 participants at the second noun in Experiment 2[2] After blink correction, we thus initially conducted analyses for a word region with only those participants that met the 27% threshold. Since results did not differ substantially, however, when including all participants, the reported analyses are those for all participants.

---

[2] For instance, a rejection rate of 5 out of 19 correctly answered trials would be 26 percent of the data for a given condition while 6 out of 19 rejected would be more than 27 percent.

Analysis of behavioral data. For response latency analyses, any score $+/-$ 2 standard deviations from the mean response latency of a participant was removed prior to further analyses and we report the original reading times. Mean response latencies, log-transformed to improve normality, and time-locked to the sentence-final word (the second noun) as well as accuracy scores, summarized by participants ($F1$) and items ($F2$), were analyzed via repeated measures ANOVAs with the role and action congruence factors (congruous vs. not)[3]. Following reliable effects in the ANOVA analyses, we conducted paired sample t-tests and we report p-values after Bonferroni (0.05/6 in Experiment 1). In Experiment 2, the selection of comparisons was guided by reliable effects in Experiment 1. For the analysis of working memory scores from the reading span test (Daneman & Carpenter, 1980), we computed the proportion of items for which a given participant recalled all the elements correctly as a proxy for VWM scores (Conway et al., 2005). For the extended complex figure test, we followed the scoring procedure for the motor-independent ECFT-MI described in Fasteneau (2003).

---

[3] The accuracy data were in addition analyzed with mixed-effects regression using a generalized linear model with a logit link function (Baayen, 2008; Bates et al., 2011; Quene & van den Bergh, 2008, lme4 package of R). Accurate responses were coded as '1', inaccurate responses as '0'. Role-relations congruence and action-congruence factors were centered prior to analyses (collin.fnc condition value = 1, indicating no issues with multi-collinearity of the predictors). We use the following models for the analysis by subjects: $lmer(accuracy \sim (1 + rolecongruence * actioncongruence|mysubj) + (1 + rolecongruence * actioncongruence|myitem) + rolecongruence * actioncongruence, data = mydata, family = binomial)$. Since these results replicated the ANOVA results, we only report the latter.

Analysis of the ERP data. Following the analysis procedure by Knoeferle et al. (2011), and based on visual inspection and traditional (sensory) evoked potential epochs, analyses of variance (ANOVAs) in Experiment 1 were conducted on the mean amplitudes of the average ERPs elicited by the first nouns (*gymnast*), the verbs (e.g., *punches*), and the second nouns (e.g., *journalist*) in three time windows each (0-100 ms, 100-300 ms, and 300-500 ms)[4]. We analyzed the first noun and early verb since we could, in principle, see early effects of the role relations mismatch. We also analyzed ERPs to the verb, where we should see a verb-action congruence effect from 300-500 ms since Knoeferle et al. (2011) reported verb-action congruence effects in this time window. Analyses of the ERPs to the second nouns (*journalist*) were motivated by previously-observed verb-action and role congruence effects. In Experiment 2, the selection of time windows and comparisons was guided by reliable effects in Experiment 1. Note that in Experiment 1 (verb) and in Experiment 2 (verb and the second noun) the standard baseline (-200 to 0 before word onset) contained reliable congruence effects. To ensure that congruence effects in the baseline did not impact the analyses for these regions, we selected a different baseline for them. All ERP analyses to the verb were baselined to -200 to 0 ms before the first noun. In Experiment 2, analyses of ERPs to both the verb and second noun were baselined to -200 to 0 ms before the first noun.

---

[4] We did not correct for the overall number of time windows for which we report analyses (10 in Experiment 1; 6 in Experiment 2; however, if we adjusted the p-values after Bonferroni (/ 10 analyses regions, adjusted p=.005), the key results and conclusions would still hold.

We performed omnibus repeated measures ANOVAs on mean ERP amplitudes (averaged by participants for each condition at each electrode site) with role congruence (incongruent vs. congruent), action 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 either role congruence (match vs. mismatch), or action congruence (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 Greenhouse-Geisser corrected p-values. In Experiment 1, we conducted six tests on mean ERP amplitudes at RMPf and at RMOc since those two sites illustrate variation of the role versus action congruence effects along the anterior-posterior dimension, Bonferroni-corrected p / 12 for 2 x 6 comparisons). In Experiment 2, we analyzed only time windows that had shown statistically significant differences in Experiment 1. For these comparisons, we conducted paired sample t-tests and we report p-values after Bonferroni.

Correlation analyses

Correlation analyses were used to ascertain to what extent end-of-sentence verification times co-varied with ERPs and the behavioral scores that we collected. Our research had revealed reliable correlations between ERP differences over right hemispheric sites at the verb and second noun phrase and sentence-final RT differences, as well as between RT differences scores, verbal working memory, and accuracy scores (Knoeferle et al., 2011).

In line with these prior analyses, we computed each participant's mean congruence effects (action mismatch minus full match and role mismatch minus full match ERP amplitude) from 300-500 ms at the first noun, verb and second noun for each of the two factors, and each participant's congruence effect for verification response latencies (action mismatch minus full match and role mismatch minus full match). Congruence ERP difference scores were averaged across the electrode sites in the four slices used for the ANOVA analyses (e.g., left lateral: LLPf, LLFr, LLTe, LDPa, LLOc). 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 response latencies, a positive number indicates longer verification times for incongruous than congruous times and a negative number indicates the converse.

For the RT-ERP correlations at the first and second nouns in the four slices of a given time window, and for correlations of ERP scores with verbal working memory scores (VWM), visual-spatial scores (ECFT), and mean accuracy, the Bonferroni correction was 0.05/4 (slices). For the RT-ERP correlation analyses at the verb we compared correlations of corresponding response time and left-lateral ERP differences (e.g., action mismatch RT with ERP differences) with correlations of response time and right-lateral ERP differences. Based on Knoeferle et al. (2011) we expect reliable correlations for action mismatch differences over the right but not left lateral slice (Bonferroni 0.05/2). Since the Kolmogorov-Smirnov test indicated normality violations for ECFT and VWM scores (Experiment 1) and for ECFT, VWM, and mean accuracy scores (Experiment 2), we report Spearman's ρ for the respective correlations (r_s). Effect sizes are reported using Cohen's d.

Results Experiment 1 (500 ms SOA)

Behavioral results

Overall accuracy was 88 % (accuracy by participants for full matches: 88 %, SD=8.30; action mismatches: 82 %, SD=9.11; role mismatches: 90 %, SD=7.30; combined mismatches: 92 %, SD=7.18). Accuracy was significantly higher for role mismatches than matches (mean difference = |5.59|, SE of the mean difference = 1.09, $F1(1, 31) = 26.18, p$ <0.001, $\eta^2 = 0.46$; $F2(1, 79) = 17.46, p < 0.001$, $\eta^2 = 0.18$) while there was no reliable accuracy difference between action matches and mismatches ($ps > 0.1$), resulting in an interaction ($F1(1, 31) = 9.24, p < 0.01, \eta^2 = 0.23$; $F2(1, 79) = 8.78, p$ <0.01, η^2=0.1). Pairwise t-tests revealed significantly less accurate responses for the action mismatch versus role mismatch condition ($t1(1, 31) = -4.48, p <$ 0.001, $d = 0.63$; $t2(1, 79) = -3.82, p < 0.01$, $d = 0.56$); for the action mismatch versus combined mismatch condition ($t1(1, 31) = -6.83, p < 0.0001$, $d = 0.78$, $t2(1, 79) = -4.81, p$ <0.0001, $d = 0.65$); and marginally more accurate responses for the full match than action mismatch condition ($t1(1, 31) =$ $2.68, p = 0.07$, $d = 0.43$; $t2(1, 79) = 2.65, p$= 0.06, $d = 0.43$, other ps > 0.6).

Response latencies were 1078 ms (SD=292.79) for full matches, 1185 ms (SD=313.23) for action mismatches, 1102 ms (SD=286.58) for role mismatches, and 1092 ms (SD=286.74) for combined mismatches (by participants). Repeated measures ANOVAS confirmed faster response times for the action matches than mismatches (1090 ms vs. 1139 ms, mean difference by subjects = |48.39|, $SE = 21.35$, $F1(1,31) = 5.02, p < 0.05, \eta^2 = 0.14$, $F2(1,79) = 1.17, p > 0.25, \eta^2 = 0.02$), but not for role matches (1131 ms) versus mismatches (1097 ms, mean difference by subjects = |33.98|, $SE = 19.84$, $F1(1,31) = 1.75, p = 0.20$, $\eta^2 = 0.05$, $F2 < 1$); the interaction between these two manipulated factors was reliable ($F1(1,31) = 7.20, p < 0.02, \eta^2 = 0.19$, $F2(1,79) = 5.01, p < 0.05$, $\eta^2 = 0.06$). Pairwise t-tests showed that responses were reliably faster by subjects for the full match versus the action mismatch condition ($t1(1,31) = -3.12, p < 0.05$, $d = 0.49$, $t2(1,79) = -2.47, p$ $= 0.1$, $d = 0.27$); marginally for the role versus action mismatch condition ($t(1,31) = 2.72, p = 0.07$, $d = 0.44$, $t2 < 1$; and for the combined versus action mismatch condition ($t(1,31) = 3.12, p < 0.05$, d=0.49, $t2 < 1.6$, other ps > 0.2). Scores for the extended complex figure test (ECFT) ranged from 8-18 with a mean of 13.25. Verbal working memory (VWM) scores ranged from 0.13 to 0.83 (mean = 0.36). These scores are comparable to previously-observed ECFT (see Fasteneau, 1999, 2003) and reading span scores (e.g., Knoeferle et al., 2011).

ERP results

Figure 1 shows grand average ERPs (N=32) at all 26 electrode sites in the four conditions time-locked to the onset of the verb. Figure 2 displays mean amplitude role mismatches versus matches, together with the spline-interpolated topographies of their difference (200-400 ms after the first noun onset, and between 300-500 ms at the second noun). Figure 3 displays the grand average ERPs (at prefrontal, parietal, temporal, and occipital sites) for action mismatch versus action matches, together with the spline-interpolated topographies of their difference (300-500 ms post-verb onset, lasting into the post-verbal determiner). Tables 2, 3, and 4 present the corresponding ANOVA results for main effects of role and action congruence and interactions between these two factors, hemisphere, laterality, and anteriority at the first noun, verb, and second noun.

Figure 1: about here

These figures and tables illustrate temporally and topographically distinct effects of role and action congruence (see Supplementary Materials II for effect sizes): During the first noun and early verb, we observed role congruence but no action congruence effects. These took the form of a somewhat anterior negativity (100-300; 200-400 ms) and an ensuing posterior positivity beginning around 400 ms after noun onset and continuing beyond the onset of the subsequent verb (0-100 ms and 100-300 ms), both larger for role mismatches than matches (Fig. 2 and Table 2). For the anterior negativity, mean amplitudes to the role mismatches (1.55 μV) were reliably more negative than to the full matches (2.82 μV) at frontal sites (RMPf, 100-300 ms: $t(1{,}31)=2.98$, $p < 0.05$, $d = 0.47$) but not occipitally (RMOc, $t < 1$, Bonferroni adjustments 0.05/12 for six tests at 2 electrode sites). Role congruence effects at the verb emerged as a broadly-distributed positivity that was descriptively somewhat larger over posterior than anterior sites (0-100 ms, see Fig. 2 and Table 3, *t*-tests for RMPf, RMOc n.s.). The role congruence positivity continued, broadly distributed, from 100-300 ms.

From 300-500 ms at the verb, role congruence effects were absent but we replicated a broadly distributed negativity (N400) that was larger for

Figure 2: about here

Figure 3: about here

action mismatches than matches; over the right than left hemisphere; over medial than lateral sites; and over posterior than anterior sites (Knoeferle et al., 2011). In contrast with the anterior role congruence negativity to the first noun, action mismatches (-1.73 μV) were more negative than the full match (0.69 μV) at RMOc ($t(1,31)$=3.54, $p < 0.02$, $d = 0.54$) but not at RMPf ($p > 0.2$), illustrating the posterior distribution; they were also more negative than the role mismatches over RMPf (-0.60 vs. 1.52 μV, t(1,31)=-3.30, $p < 0.02$, $d = 0.51$) and marginally over RMOc (-1.73 vs. 0.32 μV, $t(1,31)$=-2.92, $p = 0.07$, $d = 0.46$). At the second noun, we failed to replicate the previously observed verb-action congruence effect, but observed a broadly distributed negativity (300-500 ms, Fig. 2) which was larger for role mismatches than matches (100-300 ms and 300-500 ms, Table 4, t-tests at RMPf and RMOc n.s.).

Table 2: about here

Table 3: about here

Table 4: about here

Correlation results

At the first noun, the lower a participant's visual-spatial test scores (ECFT), the larger was her role congruence effect (Table 1 in Supplementary Materials IV, other correlations *n.s.*). Descriptively, the relationship between ERP mean amplitude differences from 300-500 ms at the verb and RT differences appears similar to the one observed by Knoeferle et al. (2011) but was not reliable ($p > 0.1$, for more details see Supplementary Materials I). At the second noun, action mismatch ERP difference scores correlated positively with action mismatch RT difference scores such that the larger a participant's mean amplitude congruence effect, the smaller her response time congruence effect and vice versa (Fig. 4a). In addition, role mismatch ERP difference scores correlated with role mismatch RT differences - the smaller the role mismatch ERP negativity, the larger the response time congruence effect (Fig. 4b). No further robust difference score correlations between ERPs and the behavioral measures were observed (see Supplementary Materials IV).

Figure 4: about here

Discussion

Role relation congruence was verified more accurately than action congruence. Moreover, role relation congruence effects preceded action congruence effects in the response times and the sentential ERPs. Our role congruence effects emerged earlier than in Wassenaar and Hagoort, namely, to the first noun (an anterior-medial negativity from 100-300 and 200-400 ms), and early in the response to the verb (a posterior positivity from 0-100 ms). By contrast, we did not observe any role relations congruence ERP effects at the verb, which did, however, show a larger N400 to action mismatches than matches. Post-verbally, role mismatches elicited a broadly-distributed larger negativity relative to the role matches. Overall, role congruence effects were distinct from, and preceded, action congruence ERP effects, implicating more than a single mismatch processor.

Why did we find earlier role congruence effects than Wassenaar and Hagoort? Some of the rapidity with which role congruence effects appeared in our study is likely due to the relatively slow word-by-word presentation (word duration was 200 ms and the SOA was 500 ms for Experiment 1). If participants have sufficient time, they may already begin to assign thematic role relations during the first noun and early verb. Wassenaar and Hagoort, by contrast, presented fluid spoken sentences (no SOA specified), and perhaps their older participants had less time between the first noun and verb to begin to process thematic role relations such that thematic role congruence effects emerged only later during the verb. Experiment 2 examines whether the key result in the RTs, ERPs, and correlations - viz. that role-relations congruence effects are distinct from and precede verb-action congruence effects - generalizes with more fluid sentence presentation.

Results Experiment 2 (300 ms SOA)

We shortened the onset asynchrony of words from 500 to 300 ms while keeping word presentation time constant (200 ms, ISI=100 ms). If the time course of the action relative to role congruence effects is invariant even at this faster presentation rate, we should replicate the observed response time, accuracy and ERP congruence effects (role congruence: noun1, 100-300 and 200-400 ms, verb: 0-100 and 100-300 ms; noun2: 100-300 and 300-500 ms; action congruence: 300-500 ms at the verb); and, if it is not invariant, we can see whether the two kinds of congruence effects vary in similar ways. Presentation rate is furthermore a parameter that existing accounts of incremental situated language processing have not explicitly included and thus a dimension along which we want to know more about visual context effects with the future goal of extending existing accounts.

Behavioral results

At 88 % the overall accuracy was comparable to Experiment 1 (by participants, full matches: 88 %, SD=10.08; action mismatches: 82 %, SD=10.55; role mismatches: 90%, SD=7.30; combined mismatches: 92 %, SD=5.95). Responses were reliably more accurate for role mismatches than matches (mean difference = $|6.16|$, $SE = 1.11$, $F1(1, 31) = 30.75, p < 0.001$, $\eta^2 = 0.50$; $F2(1, 79) = 17.50, p <$0.001, η^2 =0.18) while there was no reliable difference in response accuracy for action mismatches versus matches ($F <$ 2.1), resulting in an interaction ($F1(1, 31) = 5.90, p < 0.03$, $\eta^2 = 0.16$, $F2(1, 79) = 9.63, p <$0.01, η^2=0.11). Planned pairwise t-tests replicated reliably less accurate responses for the action mismatch versus role mismatch condition ($t1(1, 31) = -4.83, p < 0.001$, $d = 0.66$, $t2(1, 79) = -3.77, p <$

0.001, $d = 0.56$), for the action mismatch versus combined mismatch condition ($t(1, 31) = -6.71, p < 0.0001$, $d = 0.77$, $t2(1, 79) = -5.10, p <0.0001$, $d = 0.68$), and by items for the action mismatch than full match condition ($p > 0.1$ by subjects; $t2(1, 79) = -2.75, p <0.05$, $d = 0.44$, Bonferroni, 0.05/3).

Analyses of verification time latencies revealed marginal main effects by subjects of action ($F1$(1,31)=4.04, $p = 0.05$, $\eta^2 = 0.12$, $F2 < 1.5$ η^2 =0.02) and of role relations ($F1$(1,31)=3.89, p =0.06, η^2= 0.11, $F2(1, 79)$=3.03, p = 0.09, η^2=0.04, full matches: 1087 ms, SD=259.46; action mismatches: 1136 ms, SD=258.38; role mismatches: 1044 ms, SD=266.82; combined mismatches: 1093 ms, SD=259.12), and no reliable interaction ($F1 < 1$, $F2(1, 79)$=1.63, p=0.21, η^2=0.02). T-tests showed that sentences in the action mismatch condition took longer to verify than in the role mismatch condition ($t1(1, 31)$=3.24, $p < 0.01$, $d = 0.50$, $t2$ <2, Bonferroni 0.05/3, other ps > 0.09). Scores for the extended complex figure test ranged from 7 to 18 (mean = 12.09); for the reading span test participants' scores ranged from 0.09 to 0.65 (mean = 0.33), replicating Experiment 1 and Knoeferle et al. (2011).

ERP results

Figure 5 shows the grand average ERPs (N=32) at all 26 electrode sites in the full match, action mismatch, role mismatch, and combined mismatch conditions time-locked to the onset of the verb. Figure 6 displays mean amplitude role mismatches versus matches at prefrontal, parietal, temporal, and occipital sites with the spline-interpolated topographies of the differences (role mismatches minus role matches) from 200-400 ms at the first noun and from 300-500 ms at the second noun. Figure 7 displays the grand average ERPs (N=32) for action mismatches versus matches at prefrontal, parietal, temporal, and occipital sites time-locked to the first noun, together with the spline-interpolated topographies of the differences (action mismatches minus action matches) between 300-500 ms post-verb onset, and between 300-500 ms at the second noun. Tables 5 to 7 present the corresponding ANOVA results.

Figure 5: about here

These figures and tables illustrate again temporally distinct effects of role and action congruence[5]. (see Supplementary Materials III for effect sizes): a negativity during the first noun (200-400 ms) larger for role mismatching than matching sentence beginnings. Role mismatches differed reliably from full matches at RMOc ($t(1,31) = 3.64, p < 0.02$, $d = 0.55$) but not at RMPf ($p > 0.1$, i.e., the reverse anteriority pattern from Experiment 1). Combined mismatches also differed reliably from the full match condition over the posterior (RMOc, $t(1,31) = 3.79, p < 0.02$, $d = 0.56$) but not anterior (RMPf, $p > 0.2$) scalp. No further comparisons were reliable (200-400 ms, $p > 0.1$).

ANOVAs for the 0-100 ms and 100-300 ms time windows at the verb

---

[5] In Experiment 2, an error occurred in the assignment of lists: there were 16 base lists, and 32 participants such that each list should have been assigned twice (as was the case in Experiment 1). Instead, two lists were assigned only once, and 2 other lists were assigned 3 instead of 2 times. Analyses that excluded data for the lists that were assigned three times and analyses for sixteen lists replicated the reported pattern.

confirmed the same main effects and interactions as for 200-400 at the first noun. Comparisons from 0-100 ms at the verb showed reliable differences for the action versus role mismatch condition (RMPf: $t(1, 31) = 3.01, p < 0.05$, $d = 0.48$; RMOc: $t(1, 31) = 3.36, p < 0.05$, $d = 0.52$); for the role mismatch versus full match condition (RMPf: $t(1, 31) = 3.42, p < 0.05$, $d = 0.52$; RMOc: $t(1, 31) = 4.61, p < 0.001$, $d = 0.64$), and for the combined mismatch relative to the full match condition (RMPf: $p > 0.2$; RMOc: $t(1, 31) = 3.60, p < 0.02$, $d = 0.54$; other ps > 0.1). From 100-300 ms, no comparisons were reliable (ps > 0.4).

Figure 6: about here

Figure 7: about here

Table 5: about here

For the 300-500 ms time window at the verb, the role relations congruence main effect was no longer reliable. Instead, a broadly distributed negativity (N400, 300-500 ms) was larger for action mismatches than matches and maximal at centro-parietal recording sites (Fig. 7, and Table 6). The full match differed reliably from the action mismatch condition (RMPf: $t(1, 31) =$

$4.65, p < 0.001$, $d = 0.64$; RMOc: $t(1, 31) = 4.03, p < 0.001$, $d = 0.59$), and from the combined mismatch condition (RMPf: $t(1, 31) = 3.88, p < 0.02$, $d = 0.57$; RMOc: $t(1, 31) = 3.79, p < 0.02$, $d = 0.56$). Action mismatches didn't differ from role mismatches (ps > 0.2), and role mismatches didn't differ reliably from full matches ($ps > 0.08$; all other ps > 0.1).

At the second noun, we observed a right-lateralized negativity (300-500 ms), larger for action mismatches than matches (Fig. 7 and Table 7). The combined mismatch differed reliably from the full match (RMPf: $t(1, 31) = 3.38, p < 0.05$ $d = 0.52$; RMOc: $t(1, 31) = 4.01, p < 0.01$, $d = 0.58$). No further tests were reliable ($ps > 0.08$).

Table 6: about here

Table 7: about here

Correlation results

At the first noun, a participant's mean accuracy correlated with both ERP and test scores: It was higher the smaller a participant's left-lateral action mismatch difference scores (300-500 ms); and the higher her visual-spatial scores. Verbal and visual working memory scores correlated such that a higher verbal working memory score coincided with higher visual spatial scores. At the verb, action mismatch difference ERPs correlated positively with mean accuracy such that a participant with a smaller left lateral action mismatch effect from 100-300 ms had higher later accuracy (see Supplementary Materials IV).

General Discussion

With the aim of refining existing accounts of visually situated language comprehension by improving our understanding of the functional mechanisms involved, we monitored ERPs as participants inspected a picture, read a sentence, and verified whether or not the two matched in certain distinct respects. On critical trials the sentence matched the picture completely, in terms of the depicted role relations but not depicted action, vice versa, or neither. We assessed, at two SOAs (500 ms and 300 ms), whether these two types of mismatches impact written language comprehension similarly by examining (a) the time courses and scalp topographies of the associated ERP effects; and (b) correlations of these ERP effects with end of sentence response time mismatch effects, with mean accuracy in the verification task, and with participants' verbal memory and visual-spatial test scores. In short, the ERP indices of action-verb and role mismatches were not the same, implicating more than a single cognitive / neural mechanism.

Different time courses and scalp topographies

The earliest ERP effects for action mismatches (vs complete matches) emerged as a greater negativity to the mismatch between 300-500 ms relative to verb onset. By contrast, the first mismatch effect for single role relation (vs. the full match) appeared earlier - at the subject noun (100-300 and 200-400 ms in Experiment 1; 200-400 in Experiment 2), as a larger negativity, and an ensuing positivity (albeit only at the long SOAs) to the mismatch. The dual mismatch ERPs generally patterned with the role mismatch at the first noun, and with the action mismatch at the verb. Post-verbally, additional role mismatch effects (at the object noun) appeared at the long SOA and additional verb-action mismatch effects appeared at the short SOA. Response analyses revealed further differences between verb action and role-relations mismatches. At the long SOA, RTs were longer for action (but not role) mismatches than matches. Moreover, regardless of SOA, role mismatches were responded to faster and more accurately than action mismatches.

These ERP mismatch effects differed not only in their timing but also in their morphology and scalp topography. Action mismatches elicited a broadly distributed negativity maximal over posterior scalp akin to a visual N400 (see also Knoeferle et al., 2011). Indeed, this N400 effect was indistinguishable from that typically elicited by lexico-semantic anomalies or low cloze probability words in sentences read for comprehension (e.g., Kutas, 1993; Kutas et al., 2006; Otten & Van Berkum, 2007; Van Berkum et al., 1999), and likely reflects semantic matching of the verb and the action. By contrast, the role relation mismatch elicited a negativity to the first noun maximal over the *anterior* scalp throughout its course at the long SOA, and in its initial (200-400ms) phase at the short SOA consistent with more pictorial-based semantic processing (Ganis et al., 1996); its terminal phase (300-450 ms) was broadly distributed. At the long SOA, there were additional role mismatch effects at the verb (100-300 ms) and at the post-verbal object noun both anteriorly (100-300 ms) and posteriorly (300-500 ms). To reiterate, the ERP indices of action-verb and role mismatches were not the same, implicating more than a single mechanism.

Different correlation pattern

These distinct ERP mismatch effects also correlated differently with our behavioral measures. At the long SOAs, the response time congruence effects correlate with action and role-relations mismatch differences only at the second noun (but with different time courses: 0-100 ms for the verb-action mismatch effect, and from 100-300 ms for the role relation mismatch effects). Visual-spatial working memory scores did not correlate with any of the action mismatch effects but did correlate with the role relations mismatch effects at the first noun. Role relations congruence effects over left lateral sites were larger the lower the visual spatial scores (long SOA). At the short SOA, high visual spatial scores further correlated with high accuracy and with high verbal working memory; and higher accuracy coincided with smaller action mismatch effects at the verb (short SOA: from 100-300 ms left lateral).

More than one cognitive / neural mechanism underlies visual context effects during sentence comprehension

Overall then the distinct morphologies, time courses, scalp topographies, and correlation patterns of the observed congruence effects would seem to implicate more than a single mechanism in visual context effects on sentence processing. The time course differences were not expected based on the literature. Based on prior results across studies, we expected to see posterior N400s to the verb for both action (Knoeferle et al., 2011) and role-relations (Wassenaar & Hagoort, 2007) mismatches. Had these expectations been borne out, we could have argued that participants wait until the verb before matching picture-based role relations with sentence-based thematic role relations.

The role congruence effects prior to the verb (at the first noun), however, suggest more immediate incremental picture-sentence processing and active interpretation of the event depictions: It seems that when participants saw a gymnast as the patient in an event depiction, they immediately assigned a patient role (or high likelihood of patienthood) to that character; but when they then encountered *the gymnast*, in sentence-initial position, they assigned an agent role to that noun phrase, as reflected in an ERP mismatch effect. This was the case even though there was no definitive mismatch at this point in this sentence and even though among the filler sentences, some initial nouns were also thematic patients. This is a hallmark of incremental processing. Moreover, analyses with block as a within-subjects factor replicated the role relations mismatch ERP effects to the first noun absent an interaction with block (Fs < 1), suggesting these early effects are not due to participant strategies.

In principle, the distinct congruence effects to action and role mismatches could reflect the same cognitive / neural mechanism activated at different points during sentence processing. If so, then these different ERP congruence effects should have the same topography; they did not. Moreover, the presence of a positivity for role relations congruence (at the long SOA) replicates Wassenaar & Hagoort (2007) and highlights the potential contribution of structural revision to role but not action congruence processing.

Overall, the pattern of correlations is also more complex than a single cognitive / neural mechanism alone can readily accommodate. As before, we find that within participants larger action mismatch effects coincide with smaller RT congruence effects (albeit at the second noun rather than at the verb, Knoeferle et al., 2011); additionally, role mismatch effects correlated with a participant's mean accuracy, and visual-spatial scores. The latter suggests that role congruence effects may rely more on pictorial processes than do verb-action congruence effects. The correlations of action congruence effects with mean accuracy at the short SOA but with the RT congruence effect at the long SOA suggest that at the short SOA action congruence processing during the first noun and the verb contribute to processing accuracy but not speed. By contrast, at the long SOA, verb-action congruence processing seems to make more of a contribution to verification speed.

Implications for models of picture-sentence processing

In summary, these results corroborate the inadequacy of 'single-mechanism' models such as the Constituent Comparison Model by Carpenter & Just (1975) and the error monitoring account (Kolk et al., 2003). Moreover, other accounts (e.g., Altmann & Kamide, 2007; Glenberg & Robertson, 1999; Kaschak & Glenberg, 2000; Knoeferle & Crocker, 2007; Taylor & Zwaan, 2008; Zwaan & Radvansky, 1998) require some adjustment to accommodate our findings. We outline requirements / desiderata for any viable model as we work through an example for the Coordinated Interplay Account, 'CIA').

The CIA (2007). Figure 8A outlines the 2007 version of the CIA (Knoeferle & Crocker, 2007), comprising three informationally and temporally dependent steps (i to i''). As participants hear a word, they access associated linguistic and world knowledge, begin to construct an interpretation, and derive expectations (*sentence interpretation*, step i). Their interpretation and expectations can then guide (visual) attention to relevant aspects of the visual context or representations thereof (*utterance-mediated attention*, step i'); visual context representations can, in turn, be linked to the linguistic input, and if relevant, influence its interpretation (*scene integration*, step i''). This account also features a working memory (WM) component which keeps track of the interpretation (int), the expectations (ant), and representations of the scene (scene). This model, could, for instance, accommodate visual attention shifts to objects (or their previous locations) in response to object names. Its mechanisms, however, do not accommodate the distinct mismatch effects, overt verification responses, the effects of processing time, or of individual differences in WM capacity that we observed in the present experiments.

Parametrizing the CIA: Verification, timing and comprehender parameters. The Coordinated Interplay Account does not model picture-sentence verification processes per se but rather the interplay of visual attention, visual cues and utterance comprehension (see Knoeferle & Crocker, 2007, for a description). However, since verification processes seem to be part and parcel of language comprehension (see Altmann & Kamide, 1999; Knoeferle et al., 2011; Singer, 2006), and since they occur during comprehension, it is reasonable to include them into the account. The functionally distinct mismatch processes observed for action and role relations mismatches could be accommodated by having distinct picture-sentence (mis)matches feed into distinct language comprehension subprocesses such as establishing reference and thematic role assignment. We can instantiate this in the CIA through indices for the representations in WM ($\text{int}_{type\ of\ process}$, Fig. 8B). However, evidence of non-additivity (at certain time points such as the second noun and verb) suggests that these processes, while distinct, can interact as comprehension proceeds. These distinct but interacting processes could be modelled through a temporally coordinated interplay of sentence processing, attention, and visual context information to which various different mismatch processes contribute, and which subserves building of the sentence interpretation. This is already instantiated in the CIA through the temporally coordinated interplay steps (i to i'') to which both action and role congruence processes could contribute. To model functional differences indexed by different ERP topographies, we propose the engagement of different neuronal assemblies, a testable prediction in models such as CIANet (Crocker et al., 2010).

Figure 8: about here

Any viable model also would need a way of temporally tracking reactions to mismatches so as to model the extended time course of congruence processing, and an overt response index to model the post-sentence verification response latency and accuracy patterns. Both can be implemented by maintaining pictorial representations in WM, indexed as discarded; in this way, pictorial representations would remain active for some time and thereby support continued reactions to mismatch throughout the sentence up to the overt verification response. In the CIA, a truth value index for the interpretation ($\text{int}_{truth\,value}$) tracks discarded, mismatching representations, and the *response index* is set to track the value of the response as ‘true’ or ‘false’ (Fig. 8B).

Parameters that index timing and a participant’s cognitive resources can then impact the time course and interaction of different picture-sentence matching processes (whereby more time and more cognitive resources imply more in-depth, and earlier picture-sentence comparison). We could model variation of congruence effects as a function of SOA and cognitive resources by allowing these parameters to modulate either the contents of WM per se, and/or the retrieval of WM content. High verbal working memory capacity at a long SOA would thus support detailed and highly active pictorial WM representations that can then be accessed faster and lead to more pronounced role congruence effects. In the revised CIA, this is instantiated through $\mathrm{WM}_{characteristics}$ where $_{characteristics}$ could take values such as 'high' or 'low', and a timing parameter $Time_i$ which tracks processing step duration (Fig. 8B).

An illustrated example. Extended in this way, we can model the combined (dual) mismatches as follows (see Fig. 9): When participants inspect an event depiction (a journalist punching a gymnast), their role assignments (e.g., of agent to journalist and patient to gymnast) are tracked in the scene representations, $\mathrm{scene}_{i''-1}$ (step i, Fig. 9).

When they subsequently read the first noun phrase in the sentence *The gymnast applauds the journalist*, role congruence ERP effects emerge (the relative negativity and positivity to the first noun). In the model, the first noun receives an agent role (int_i [$\mathrm{GYMNAST}_{AG}$], step_i) and is indexed to the role filler (the gymnast, depicted as a patient), yielding a corresponding role relations mismatch (co-indexing, at $\mathrm{step}_{i''}$). After co-indexing, the interpretation $\mathrm{int}_{i''}$ for the long SOA would contain an agent role representation [$\mathrm{GYMNAST}_{AG-RR-M}$], where 'RR-M' specifies the role relations mismatch;

working memory would further contain a (discarded) visual representation of the first noun's referent as a patient (scene$_{i''}$[GYMNAST$_{PAT-RR-M}$]); the representation of a punching action (scene$_{i''}$[PUNCHING$_V$]), and of the journalist as the agent (scene$_{i''}$ [JOURNALIST$_{AG}$])); the response index would be set to [false]. At the short SOA, participants have less time to access the contents of working memory, possibly leading to less in-depth role congruence processing at the first noun, perhaps explaining the absence of the posterior positivity that was present at the longer SOA.

Figure 9: about here

At the verb, a verb-action congruence N400 emerges for the combined mismatches. In the model, the verb (word $i + 1$), is indexed to the action, which likewise fails. Once the verb and action have been co-indexed at step$_{i''+1}$, the interpretation thus would contain an agent noun phrase [GYMNAST$_{AG-RR-M}$], the sentential verb ([APPLAUDS$_{VA-M}$]), both marked as mismatches, and working memory also would contain a (discarded) visual representation of the first noun phrase referent (GYMNAST$_{PAT-RR-M}$), a (discarded) representation of the mismatching action (PUNCHING$_{VA-M}$), as well as the representation of a journalist in an agent role (JOURNALIST$_{AG}$, step$_{i''+1}$, Fig. 10).

At the post-verbal noun, action congruence effects were absent for the long SOA but role congruence effects were in evidence. This could be accommodated through (re-activation of) mismatching role relations representations at the post-verbal noun since that noun is implicated in thematic role assignment. No such incongruence would be expected based on this mechanism for action congruence (though note that we have observed punctate action congruence effects at the long SOA previously, Knoeferle et al., 2011). At the short SOA, by contrast, no role congruence effects emerged post-verbally and action congruence effects lasted into the post-verbal noun phrase. This could be the result of the greater compactness of word presentation (i.e., relative to the long SOA, the post-verbal noun phrase appears earlier and its presentation thus overlaps with the verb-action congruence effect) and less time to re-access role representations at the post-verbal noun at the short SOA.

Figure 10: about here

At sentence end, the response must be executed. Working memory at this point would contain the interpretation, mismatching representations, and an index of the to-be-executed response (here: ‘false’). RT action congruence effects only emerged at the long SOA. We thus speculate that at the short SOA, with less time at each word, processing was relatively more shallow, perhaps due to a good-enough strategy for representation building (e.g., Ferreira et al., 2002), or because the shorter sentence duration in combination with the pressure to respond precluded renewed access to existing WM representations for the mismatches. The absence of mismatch RT effects to role relations incongruence at the long SOA could come from processing that starts earlier for role (vs. action) mismatches, and is completed by the time the response is given such that working memory no longer contains the discarded mismatching representations. This is supported by faster response times to role than to action mismatches and by the presence of marginal role congruence effects in RTs at the short SOA.

In sum, these data place some important constraints on the specifications of viable models of situated language processing. Accounts that postulate a single representational substrate for different aspects of language and visual context, or a single mechanism, won't suffice. In general terms, it would appear then that a model in which "utterance meaning, scene information, and linguistic expectation are representationally indistinguishable and reside within a unitary system that learns, represents, and processes language and the world" would fail to explain our results (Altmann & Mirković, 2009, p. 593). They are, by contrast, compatible with models that postulate a rapid interaction between linguistic and non-linguistic information (e.g., Anderson et al., 2011; Tanenhaus et al., 1995).

We have outlined an instantiation of the CIA in which distinct comprehension (sub)processes recruit relevant aspects of a visual scene. This instantiation makes no predictions about the potential contribution of sensorimotor activations as in various embodiment accounts; however, this might be assessed by monitoring which brain regions are activated by the different mismatches (e.g. action information might be associated with sensorimotor activation; role information with visual activation). Yet another possibility is that our findings reflect a mismatch between the sentence and a subvocal verbal representation of the scene. This possibility could be tested by precluding subvocalization via engagement in an unrelated vocalisation task during scene inspection. To the extent that we obtain the same pattern of effects (e.g., the earlier effects of role-relation than verb-action mismatches), we would conclude that the underlying representations are visual rather than verbal in nature. While these are interesting avenues for future research, the contribution of the current work is a principled account of how visual context affects functionally distinct comprehension processes (during situated comprehension), especially as inferred from ERPs to various sorts of picture-sentence (mis)matches.

Tenable accounts must permit interactions between different (verb-action and role-relations) picture-sentence matching processes while keeping them distinct, representationally or otherwise. Additionally, these accounts must further specify processing parameters (time) and comprehender parameters (e.g., cognitive capacities) since at least the few parameters that we tested had distinct influences on the different picture-sentence matching processes. The proposed revised CIA accomplishes just that.

Acknowledgments

This research was funded by a postdoctoral fellowship to PK (German research foundation, DFG), the Cognitive Interaction Technology Excellence Center (DFG), and by NIH grants HD-22614 and AG-08313 to Marta Kutas. The studies were conducted while PK was at UC San Diego.

References

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

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*, 502–518.

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

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*, 583–609.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, *137*, 181–189.

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: CUP.

Barsalou, L. W. (1999). Perceptual and symbol systems. *Behavioural and Brain Sciences*, *22*, 577–609.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects* *models using S4 classes. R package version 0.999375-41. http:// CRAN.R-Project.org/package=lme4*.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Carlson, G. N., & Tanenhaus, M. K. (1989). Syntax and semantics volume 21: Thematic relations. chapter Thematic Roles and Language Comprehension. (pp. 263–287). San Diego, CA: Academic Press.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: a psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73.

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

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.

Crocker, M. W., Knoeferle, P., & Mayberry, M. (2010). Situated sentence comprehension: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, *112*, 189–201.

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

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory*, *6*, 122–149.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Fasteneau, P. (2003). Extended complex figure test. Los Angeles, CA.

Fasteneau, P. S. (1999). Development and preliminary standardization of the "Extended Complex Figure Test" (ECFT). *Journal of Clinical and Experimental Neuropsychology*, *18*, 63–76.

Federmeier, K. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491–505.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15.

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for "common sense": an electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience*, *8*, 89–106.

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

Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, *5*, 107–111.

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, *15*, 883–899.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*, 101–122.

Kamide, Y., Altmann, G. T. M., & Haywood, S. (2003). The time course of prediction in incremental sentence processing: evidence from anticipatory eye-movements. *Journal of Memory and Language*, *49*, 133–156.

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

Knoeferle, P., & Crocker, M. W. (2005). Incremental effects of mismatch during picture-sentence integration: Evidence from eye-tracking. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Conference* (pp. 1166–1171). Mahwah, NJ: Erlbaum.

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*, *75*, 519–543.

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., Urbach, T. P., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence comprehension: insights from erps and picture-sentence verification. *Psychophysiology*, *48*, 495–506.

Kolk, H., Chwilla, D., Herten, M. V., & Oor, P. (2003). Structure and limited capacity in verbal working memory: a study with event-related potentials. *Brain and Language*, *85*, 1–36.

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*, 99–116.

Kutas, M., Van Petten, C., & Kluender, R. (2006). Psycholinguistics electrified ii: 1994-2005. In M. Traxler, & M. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (pp. 659–724). New York: Elsevier. (2nd ed.).

Levy, R. (2008). Expectations-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Otten, M., & Van Berkum, J. (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.

Pickering, M., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*, 105–110.

Quene, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

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.

Sternberg, S. (1969). The discovery of processing stages: extensions of donders' method. *Acta Psychologica*, *30*, 276–315.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 632–634.

Taylor, L. J., & Zwaan, R. A. (2008). Motor resonance and linguistic focus. *Quarterly Journal of Experimental Psychology*, *61*, 896–904.

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

Van de Meerendonk, N., Kolk, H., Chwilla, D., & Vissers, C. T. (2009). Monitoring in language perception. *Language and Linguistics Compass*, *3*, 1211–1224.

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*, (pp. 967–982.).

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

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

Figure captions

Figure 1: Grand average ERPs (mean amplitude) for all 26 electrodes, right-lateral, left-lateral, right-horizontal, left-horizontal eye electrodes (‘rle’, ‘lle’, ‘rhz’ and ‘lhz’), and the mastoid (‘A2’) time-locked to the verb (Experiment 1). Negative is plotted up in all time course figures, and 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.

Figure 2: Grand average mean amplitude ERPs for role mismatching conditions versus role matching conditions across the sentence at prefrontal, parietal, temporal, and occipital sites together with the spline interpolated maps of the difference waves at the first noun (200-400 ms) and second noun (300-500 ms) in Experiment 1

Figure 3: Grand average mean amplitude ERPs for action mismatching versus matching conditions at prefrontal, parietal, temporal, and occipital sites (Experiment 1). Spline interpolated maps of the scalp potential distributions show the verb N400 (300-500 ms). In these and subsequent figures, each isopotential contour spans 0.625 μV. More negative potentials have darker shades and more positive potentials lighter shades.

Figure 4: Correlations at the second noun in Experiment 1: (a) RT and ERP action mismatch difference scores (0-100 ms); (b) RT and ERP role

mismatch difference scores (100-300 ms)

Figure 5: 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 (Experiment 2). 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.

Figure 6: Grand average mean amplitude ERPs scores for action mismatches versus matches across the sentence at prefrontal, parietal, temporal, and occipital sites (Experiment 2). Spline interpolated maps of the scalp potential distributions show the verb-action congruence N400 from 300-500 ms at the verb and from 300-500 ms at the second noun. Note that in this figure and Figure 6 the scalp potential distributions at the second noun were computed relative to a -200 to 0 baseline of the first noun

Figure 7: Grand average mean amplitude ERPs scores for role mismatches versus matches across the sentence at prefrontal, parietal, temporal, and occipital sites (Experiment 2)

Figure 8: A: The 2007 version of the Coordinated Interplay. B: the revised version of the CIA

Figure 9: State of the model when the phrase *The gymnast* is processed

after participants have seen a picture of a journalist punching a gymnast. Steps i to i'' represent the three processing stages for *the gymnast*. Step $i+1$ indicates the continuation of the loop to the next word.

Figure 10: State of the model when the verb *applauds* is processed after participants have seen a picture of a journalist punching a gymnast. Steps $i+1$ to $i''+1$ represent the three processing stages for *applauds*. Step $i+2$ indicates the continuation of the loop to the next phrase, *the journalist*.

Table 1

Table 1: Example of the four experimental conditions

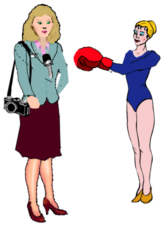
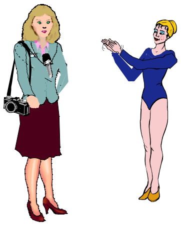
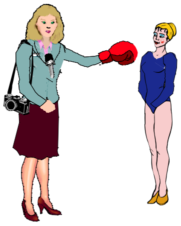
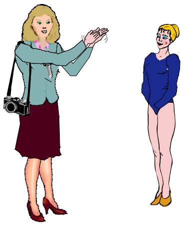
Condition	Picture		Sentence
full match	1a		*The gymnast punches the journalist*
action mismatch	1b		*The gymnast punches the journalist*
role mismatch	1c		*The gymnast punches the journalist*
combined mismatch	1d		*The gymnast punches the journalist*

Table 2: ANOVA results for first noun in Experiment 1 (SOA: 500 ms). 'R(ole)'=Role relations congruence factor; 'V(action)'=Verb-action congruence factor; Columns 4-5 show the results of the overall ANOVA electrode sets at the verb (20 electrode sites), all other $p-$values involving the independent variables in these time windows > 0.07; columns 6-9 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- and $p-$values; we report main effects of role congruence (R(ole)), action congruence (V(Action)), and interactions of these two factors with hemisphere (H), laterality (L), and anteriority (A); main effects of factors hemisphere, laterality, and anteriority are omitted for the sake of brevity as are interactions between just these three factors; degrees of freedom df(1,31) expect for RA, VA, RVA, RHA, VHA, RLA, VLA, RVHA, RVLA, RHLA, VHLA, RVHLA, df(4,124). ? $0.07 > p > 0.05$; *$p < 0.05$; **$p < .01$; ***$p < .001$;

Sentence position	Time window	Factors	Overall ANOVA	p-value	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
Noun1	0-100	–						
	100-300	Role	4.88	0.035*	1.49	4.15?	4.00?	6.17*
		RL	4.58	0.040*				
		RLA	2.66	0.055?				
	200-400	Role	9.69	0.004**	5.25*	8.78**	8.45**	10.58**
		RL	6.91	0.013*				
	300-500	RL	4.10	0.052?				

Table 3

Table 3: ANOVA results for the verb in Experiment 1 (SOA: 500 ms, baselined to 0-200 ms prior to the first noun). All other $p-$values involving the independent variables in these time windows > 0.07. ? $0.07 > p > 0.05$; $*p < 0.05$; $**p < .01$; $***p < .001$;

Sentence position	Time window	Factors	Overall ANOVA	*p*-value	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
Verb	0-100 ms	RA	4.21	0.036*	4.89*	2.54	3.96*	3.91*
		RLA	3.67	0.015*				
		RVHL	4.83	0.036*				
	100-300 ms	Role	6.13	0.019*	4.83*	3.38?	9.96**	4.23*
	300-500 ms	VAction	16.05	0.000***	9.78**	12.70**	16.78***	15.80***
		VH	8.07	0.008**				
		VL	4.63	0.039*				
		VA	4.82	0.019 ***				

Table 4

Table 4: ANOVA results for the second noun in Experiment 1 (SOA: 500 ms). All other $p-$values involving the independent variables in these time windows $> 0.07. * p < 0.05; ** p < .01; *** p < .001;$

Sentence position	Time window	Factors	Overall ANOVA	*p*-value	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
Noun2	0-100	RV	3.74	0.062?				
	100-300	RA	6.39	0.002**				
	300-500	Role	4.99	0.033*	3.48?	3.81?	7.95**	3.24?
		RVA	4.76	0.018*				

Table 5

Table 5: ANOVA results for the first noun in Experiment 2 (SOA: 300 ms). All other $p-$values involving the independent variables in these time windows > 0.07.

Sentence position	Time window	Factors	Overall ANOVA	p-value	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
Noun1	100-300	RHA	3.30	0.030*	-	-	-	
		VHA	3.09	0.035*	-	-	-	-
	200-400	Role	12.14	0.001**	6.79*	12.87**	10.91**	11.63**
		RL	10.06	0.003**				
		RHA	2.85	0.05?				
		RLA	4.71	0.004**				

Table 6: ANOVA results for the verb in Experiment 2 (SOA: 300 ms). Analyses for the verb were conducted with a 200 ms baseline prior to the first noun. All other $p-$values involving the independent variables in these time windows > 0.07.

Sentence position	Time window	Factors	Overall ANOVA	p-value	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
Verb	0-100	Role	22.04	0.000***	11.98**	22.19***	20.89***	20.67***
		RL	15.37	0.000***				
		RHA	2.63	0.063?				
		RLA	5.26	0.002**				
	100-300	RL	10.22	0.003**				
		RLA	3.69	0.017*				
	300-500	VAction	17.42	0.000***	4.08?	15.87***	18.54***	24.86***
		RV	12.87	0.001**				
		VH	9.88	0.004**				
		RL	19.03	0.000***				
		VL	29.22	0.000***				
		RVL	10.71	0.003**				
		VHL	4.81	0.036*				
		VHA	3.67	0.028*				
		RLA	4.19	0.007**				
		VLA	2.91	0.040*				
		RVLA	2.96	0.036*				
		VHLA	3.32	0.031*				
		RVHLA	2.97	0.030*				

Table 7: ANOVA results for the second noun in Experiment 2 (SOA: 300 ms). All other $p-$values involving the independent variables in these time windows > 0.05.

Sentence position	Time window	Factors	Overall ANOVA	p-value	Left lateral sites	Left medial sites	Right lateral sites	Right medial sites
Noun 2	300-500	VAction	7.11	0.012*	0.30	8.82**	6.38*	11.22**
		VH	7.86	0.009**				
		RL	11.51	0.002**				
		VL	20.55	0.000***				
		VHL	9.54	0.004**				
		RLA	4.15	0.005**				
		VLA	2.96	0.030*				
		VHLA	5.42	0.002**				

Figure 1

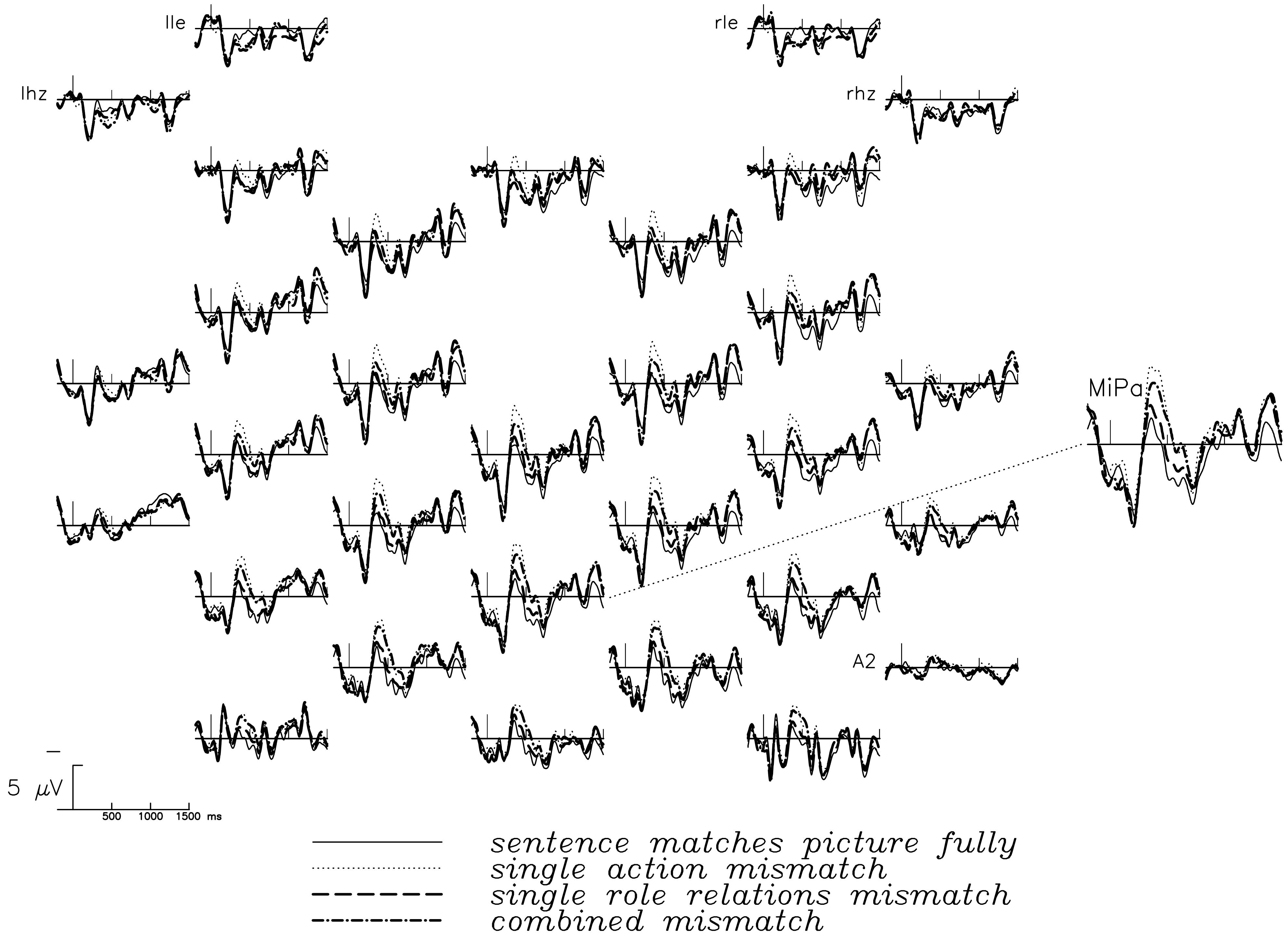

lle
rle
lhz
rhz
MiPa
A2
5 μV
500
1000
1500
ms
sentence matches picture fully
single action mismatch
single role relations mismatch
combined mismatch

Figure 2

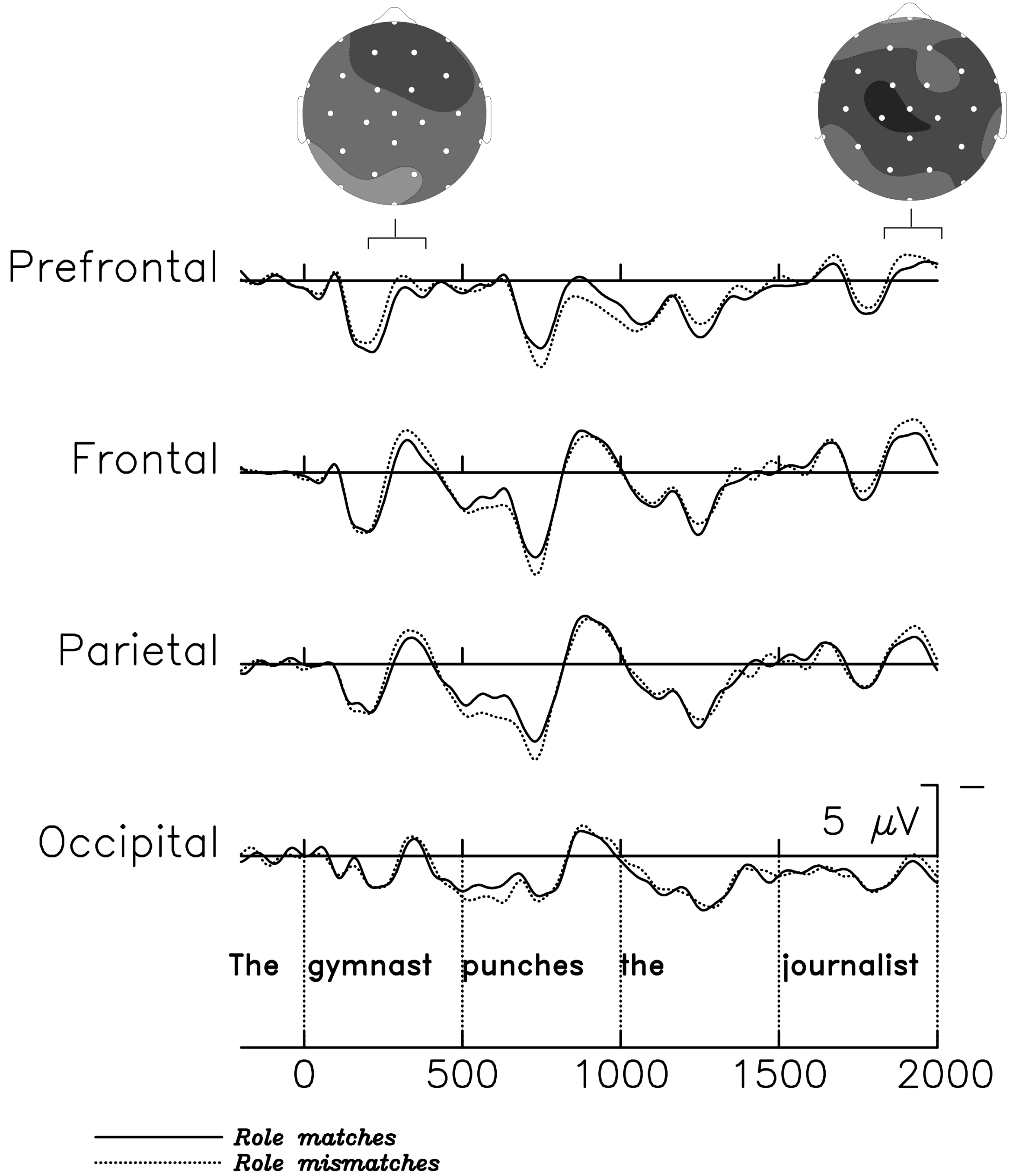
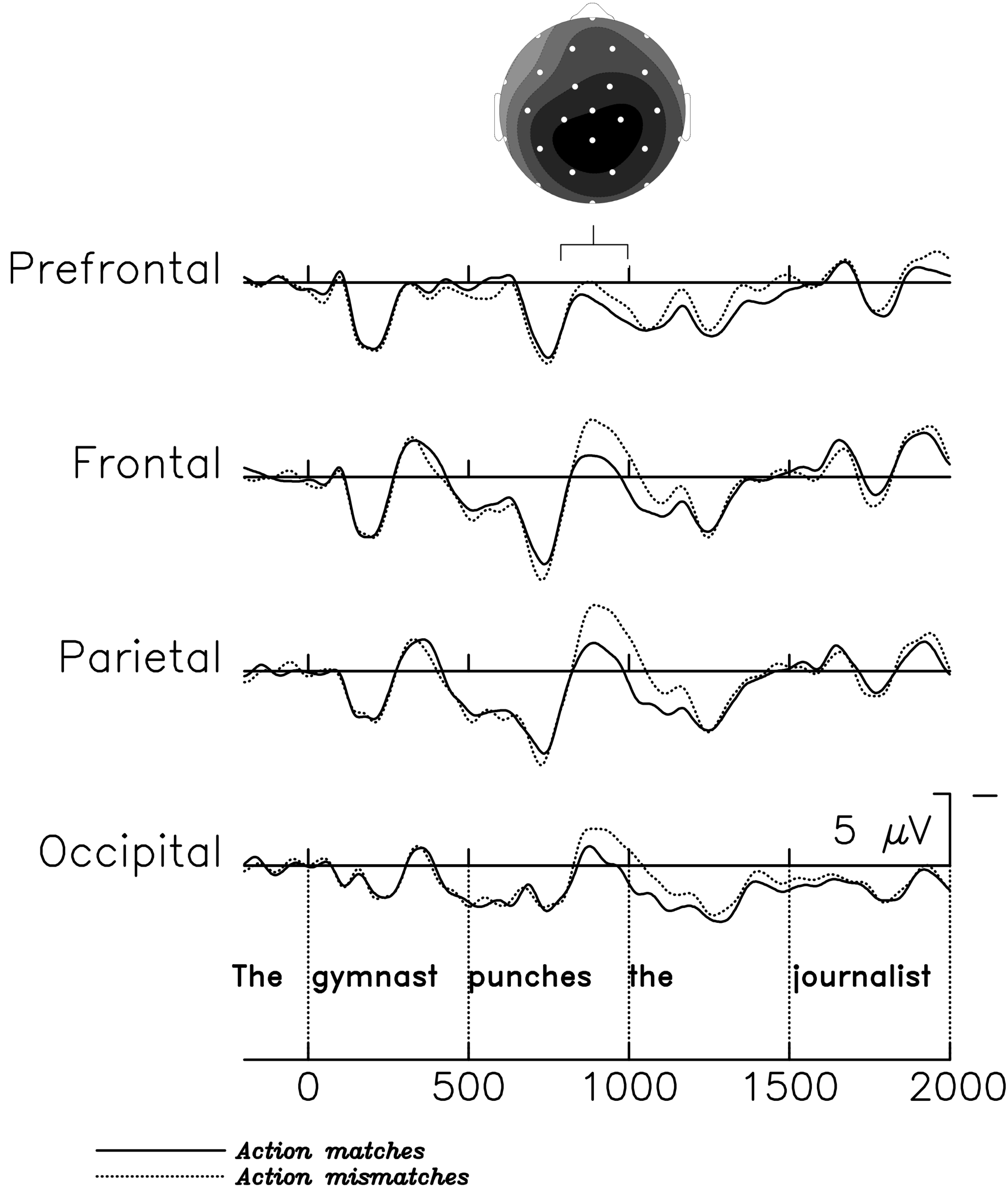
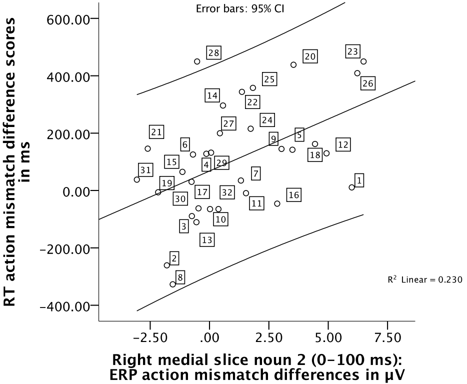
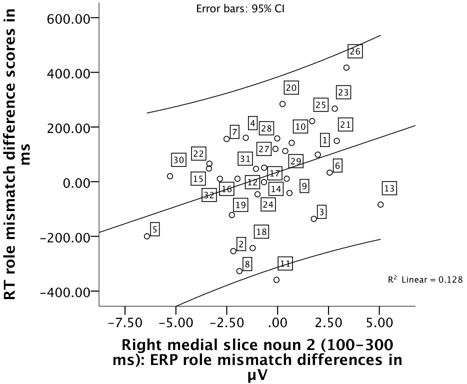

Prefrontal
Frontal
Parietal
Occipital
5 μV
–
The
gymnast
punches
the
journalist
0
500
1000
1500
2000
Role matches
Role mismatches

Figure 3


Prefrontal
Frontal
Parietal
Occipital
5 μV
–
The
gymnast
punches
the
journalist
0
500
1000
1500
2000
Action matches
Action mismatches


Error bars: 95% CI
RT action mismatch difference scores in ms
600.00
400.00
200.00
0.00
−200.00
−400.00
−2.50
.00
2.50
5.00
7.50
Right medial slice noun 2 (0–100 ms): ERP action mismatch differences in μV
R² Linear = 0.230
28
20
23
26
25
14
22
24
27
21
6
9
5
12
18
15
4
29
31
7
19
1
30
17
32
16
11
10
3
13
2
8

(a)


Error bars: 95% CI
RT role mismatch difference scores in ms
600.00
400.00
200.00
0.00
−200.00
−400.00
−7.50
−5.00
−2.50
.00
2.50
5.00
Right medial slice noun 2 (100–300 ms): ERP role mismatch differences in μV
R² Linear = 0.128
26
20
23
25
4
28
10
21
7
1
27
22
30
31
29
6
17
15
12
9
16
14
13
32
19
24
3
5
18
2
8
11

(b)

Figure 5

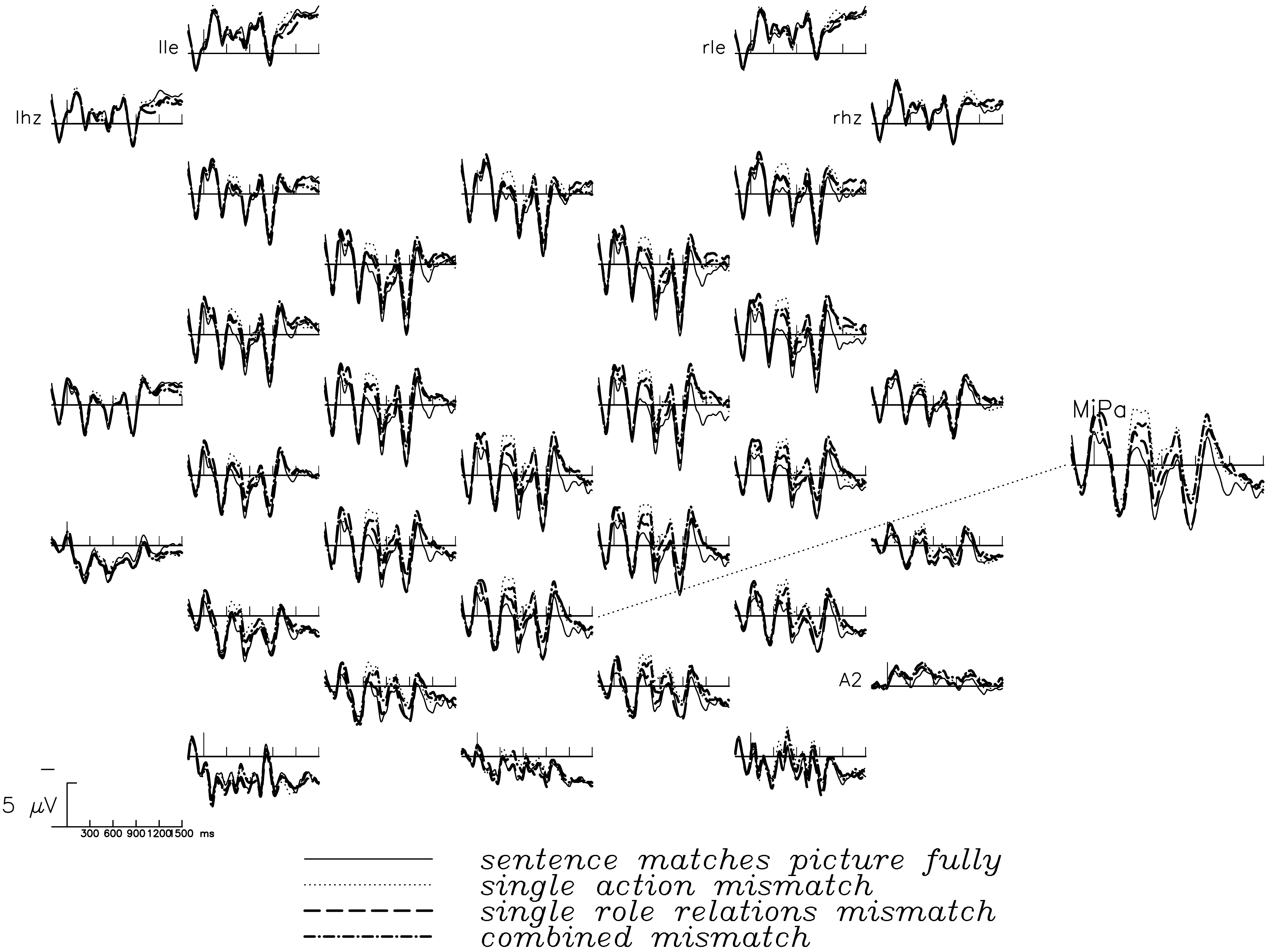

lle
rle
lhz
rhz
MiPa
A2
5 μV
300 600 900 1200 1500 ms
sentence matches picture fully
single action mismatch
single role relations mismatch
combined mismatch

Figure 6

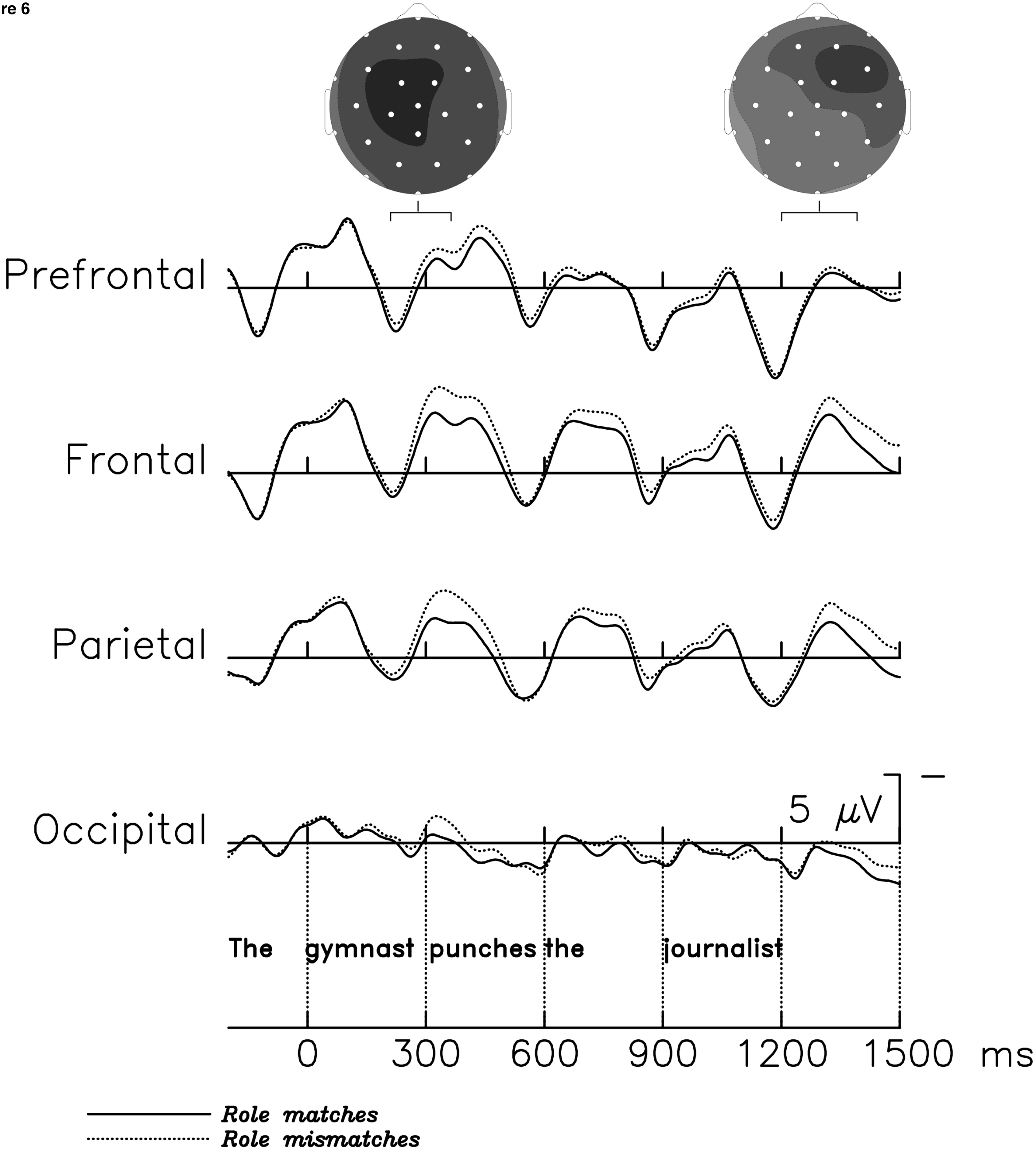

Prefrontal
Frontal
Parietal
Occipital
5 μV
−
The
gymnast
punches
the
journalist
0
300
600
900
1200
1500
ms
Role matches
Role mismatches

Figure 7

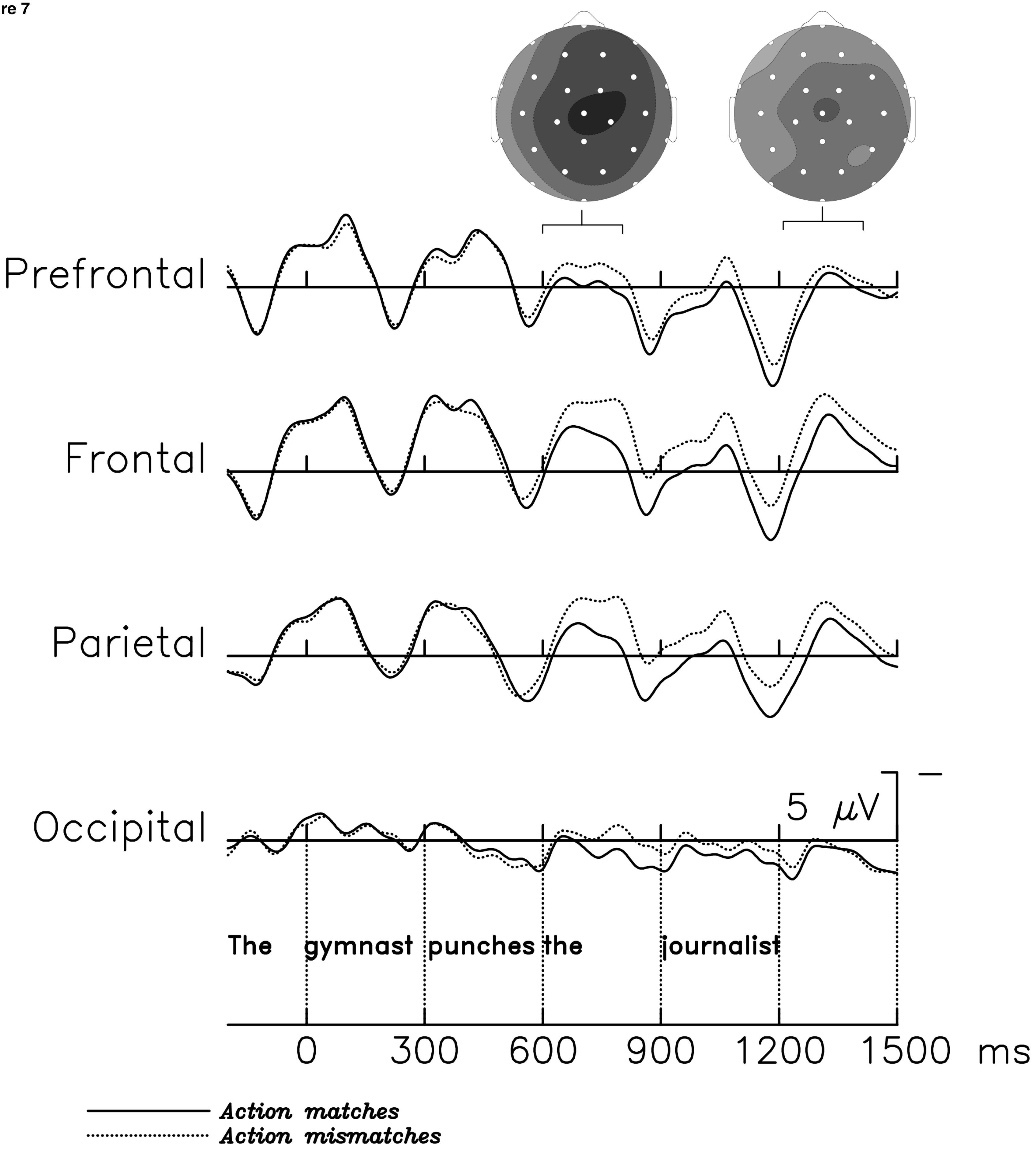

Prefrontal
Frontal
Parietal
Occipital
5 μV
–
The
gymnast
punches
the
journalist
0
300
600
900
1200
1500
ms
Action matches
Action mismatches

Figure 8

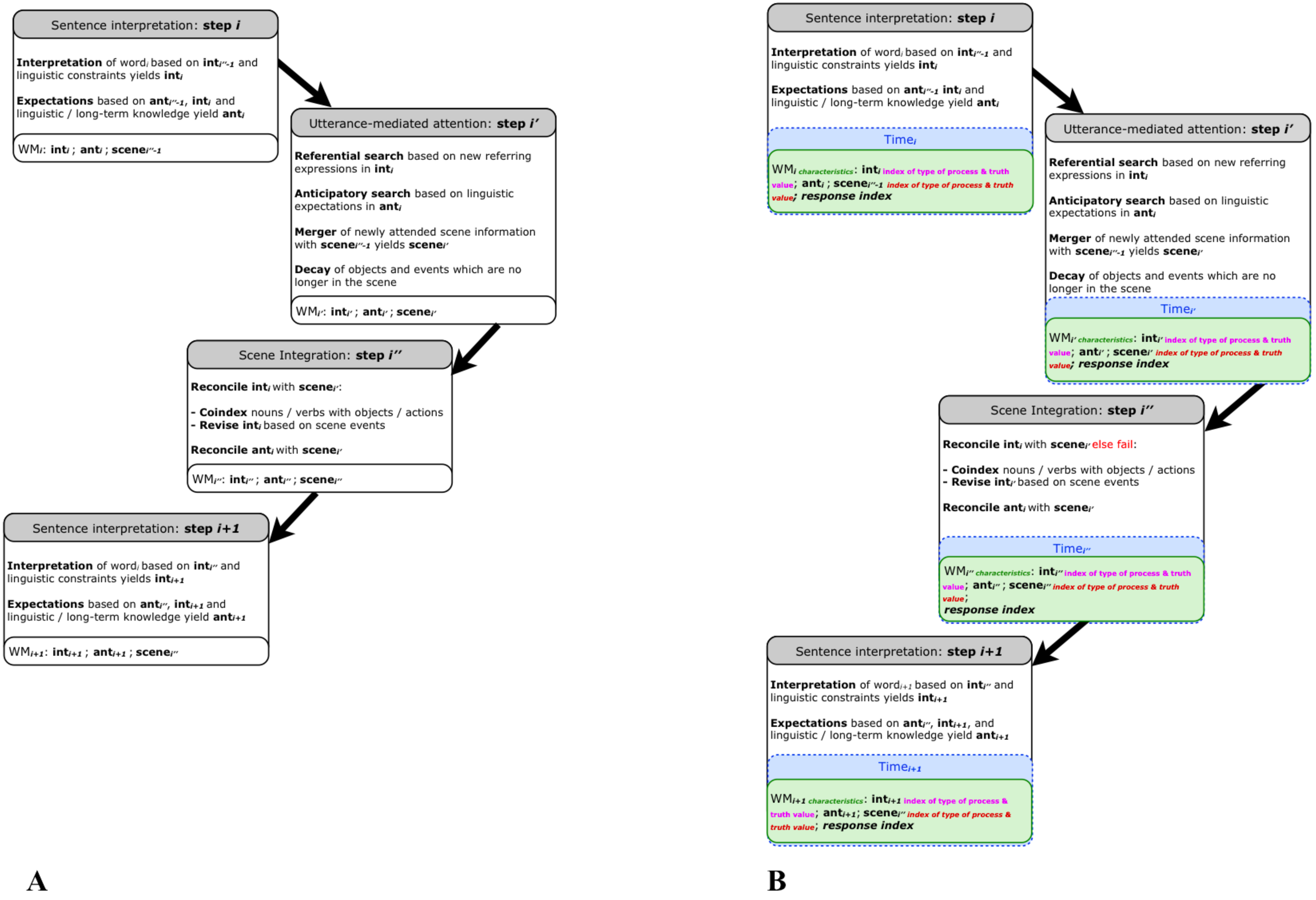
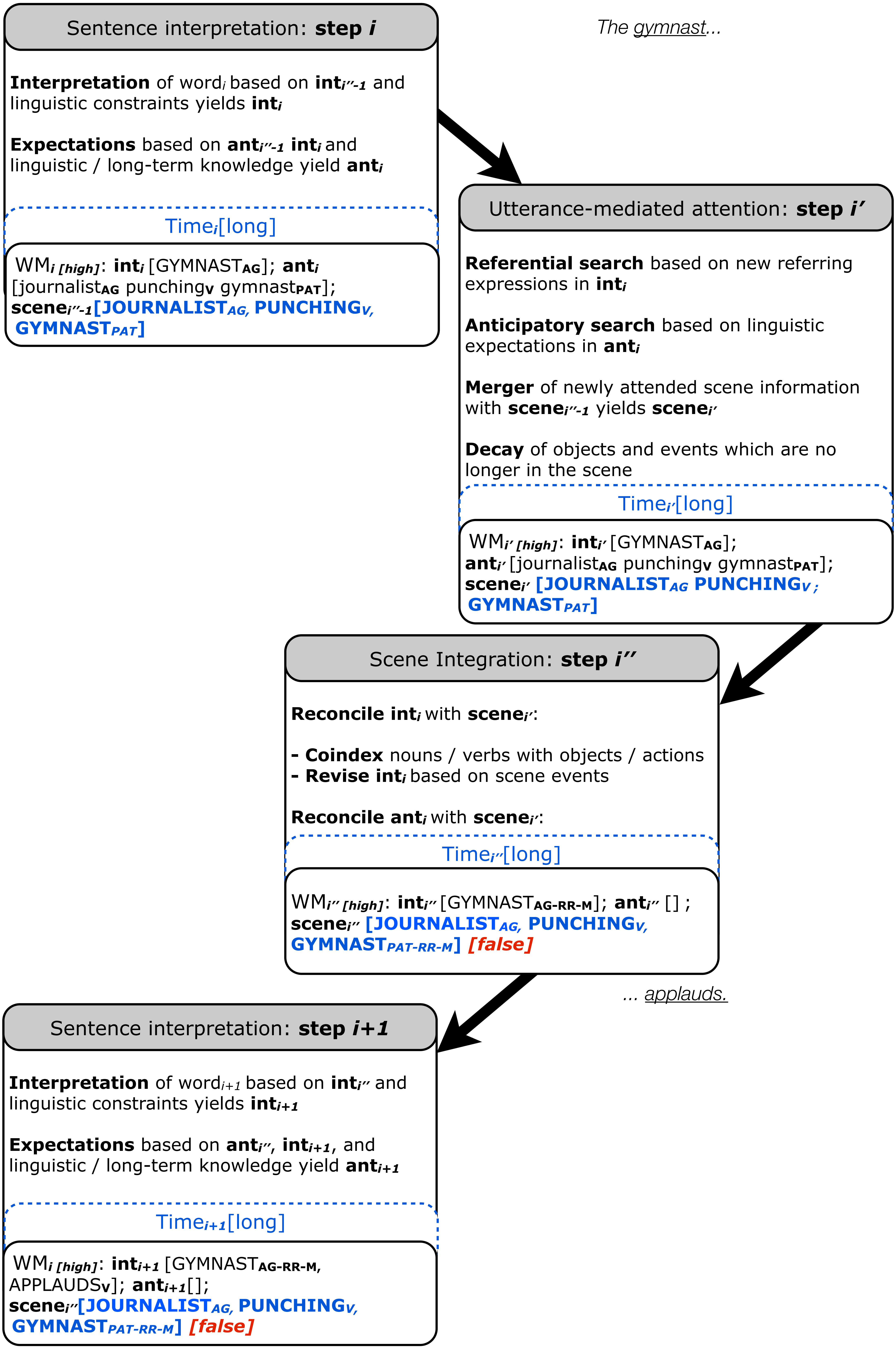

Sentence interpretation: **step *i***
Interpretation of word$_i$ based on **int**$_{i''-1}$ and linguistic constraints yields **int**$_i$
Expectations based on **ant**$_{i''-1}$, **int**$_i$ and linguistic / long-term knowledge yield **ant**$_i$
WM$_i$: **int**$_i$; **ant**$_i$; **scene**$_{i''-1}$
Utterance-mediated attention: **step *i'***
Referential search based on new referring expressions in **int**$_i$
Anticipatory search based on linguistic expectations in **ant**$_i$
Merger of newly attended scene information with **scene**$_{i''-1}$ yields **scene**$_{i'}$
Decay of objects and events which are no longer in the scene
WM$_{i'}$: **int**$_{i'}$; **ant**$_{i'}$; **scene**$_{i'}$
Scene Integration: **step *i''***
Reconcile int$_i$ with **scene**$_{i'}$:
- **Coindex** nouns / verbs with objects / actions
- **Revise int**$_i$ based on scene events
Reconcile ant$_i$ with **scene**$_{i'}$
WM$_{i''}$: **int**$_{i''}$; **ant**$_{i''}$; **scene**$_{i''}$
Sentence interpretation: **step *i+1***
Interpretation of word$_i$ based on **int**$_{i''}$ and linguistic constraints yields **int**$_{i+1}$
Expectations based on **ant**$_{i''}$, **int**$_{i+1}$ and linguistic / long-term knowledge yield **ant**$_{i+1}$
WM$_{i+1}$: **int**$_{i+1}$; **ant**$_{i+1}$; **scene**$_{i''}$
A
Sentence interpretation: **step *i***
Interpretation of word$_i$ based on **int**$_{i''-1}$ and linguistic constraints yields **int**$_i$
Expectations based on **ant**$_{i''-1}$ **int**$_i$ and linguistic / long-term knowledge yield **ant**$_i$
Time$_i$
WM$_i$ *characteristics*: **int**$_i$ *index of type of process & truth value*; **ant**$_i$; **scene**$_{i''-1}$ *index of type of process & truth value*; ***response index***
Utterance-mediated attention: **step *i'***
Referential search based on new referring expressions in **int**$_i$
Anticipatory search based on linguistic expectations in **ant**$_i$
Merger of newly attended scene information with **scene**$_{i''-1}$ yields **scene**$_{i'}$
Decay of objects and events which are no longer in the scene
Time$_{i'}$
WM$_{i'}$ *characteristics*: **int**$_{i'}$ *index of type of process & truth value*; **ant**$_{i'}$; **scene**$_{i'}$ *index of type of process & truth value*; ***response index***
Scene Integration: **step *i''***
Reconcile int$_i$ with **scene**$_{i'}$ else fail:
- **Coindex** nouns / verbs with objects / actions
- **Revise int**$_{i'}$ based on scene events
Reconcile ant$_i$ with **scene**$_{i'}$
Time$_{i''}$
WM$_{i''}$ *characteristics*: **int**$_{i''}$ *index of type of process & truth value*; **ant**$_{i''}$; **scene**$_{i''}$ *index of type of process & truth value*;
response index
Sentence interpretation: **step *i+1***
Interpretation of word$_{i+1}$ based on **int**$_{i''}$ and linguistic constraints yields **int**$_{i+1}$
Expectations based on **ant**$_{i''}$, **int**$_{i+1}$, and linguistic / long-term knowledge yield **ant**$_{i+1}$
Time$_{i+1}$
WM$_{i+1}$ *characteristics*: **int**$_{i+1}$ *index of type of process & truth value*; **ant**$_{i+1}$; **scene**$_{i''}$ *index of type of process & truth value*; ***response index***
B

Figure 9



The gymnast...

Sentence interpretation: step *i*

Interpretation of word$_i$ based on **int**$_{i''-1}$ and linguistic constraints yields **int**$_i$

Expectations based on **ant**$_{i''-1}$ **int**$_i$ and linguistic / long-term knowledge yield **ant**$_i$

Time$_i$[long]

WM$_{i\,[high]}$: **int**$_i$ [GYMNAST$_{AG}$]; **ant**$_i$ [journalist$_{AG}$ punching$_V$ gymnast$_{PAT}$]; **scene**$_{i''-1}$**[JOURNALIST$_{AG}$, PUNCHING$_V$, GYMNAST$_{PAT}$]**

Utterance-mediated attention: step *i′*

Referential search based on new referring expressions in **int**$_i$

Anticipatory search based on linguistic expectations in **ant**$_i$

Merger of newly attended scene information with **scene**$_{i''-1}$ yields **scene**$_{i'}$

Decay of objects and events which are no longer in the scene

Time$_{i'}$[long]

WM$_{i'\,[high]}$: **int**$_{i'}$ [GYMNAST$_{AG}$]; **ant**$_{i'}$ [journalist$_{AG}$ punching$_V$ gymnast$_{PAT}$]; **scene**$_{i'}$ **[JOURNALIST$_{AG}$ PUNCHING$_V$; GYMNAST$_{PAT}$]**

Scene Integration: step *i″*

Reconcile int$_i$ with **scene**$_{i'}$:

- **Coindex** nouns / verbs with objects / actions
- **Revise int**$_i$ based on scene events

Reconcile ant$_i$ with **scene**$_{i'}$:

Time$_{i''}$[long]

WM$_{i''\,[high]}$: **int**$_{i''}$ [GYMNAST$_{AG\text{-}RR\text{-}M}$]; **ant**$_{i''}$ [] ; **scene**$_{i''}$ **[JOURNALIST$_{AG}$, PUNCHING$_V$, GYMNAST$_{PAT\text{-}RR\text{-}M}$]** ***[false]***

... applauds.

Sentence interpretation: step *i+1*

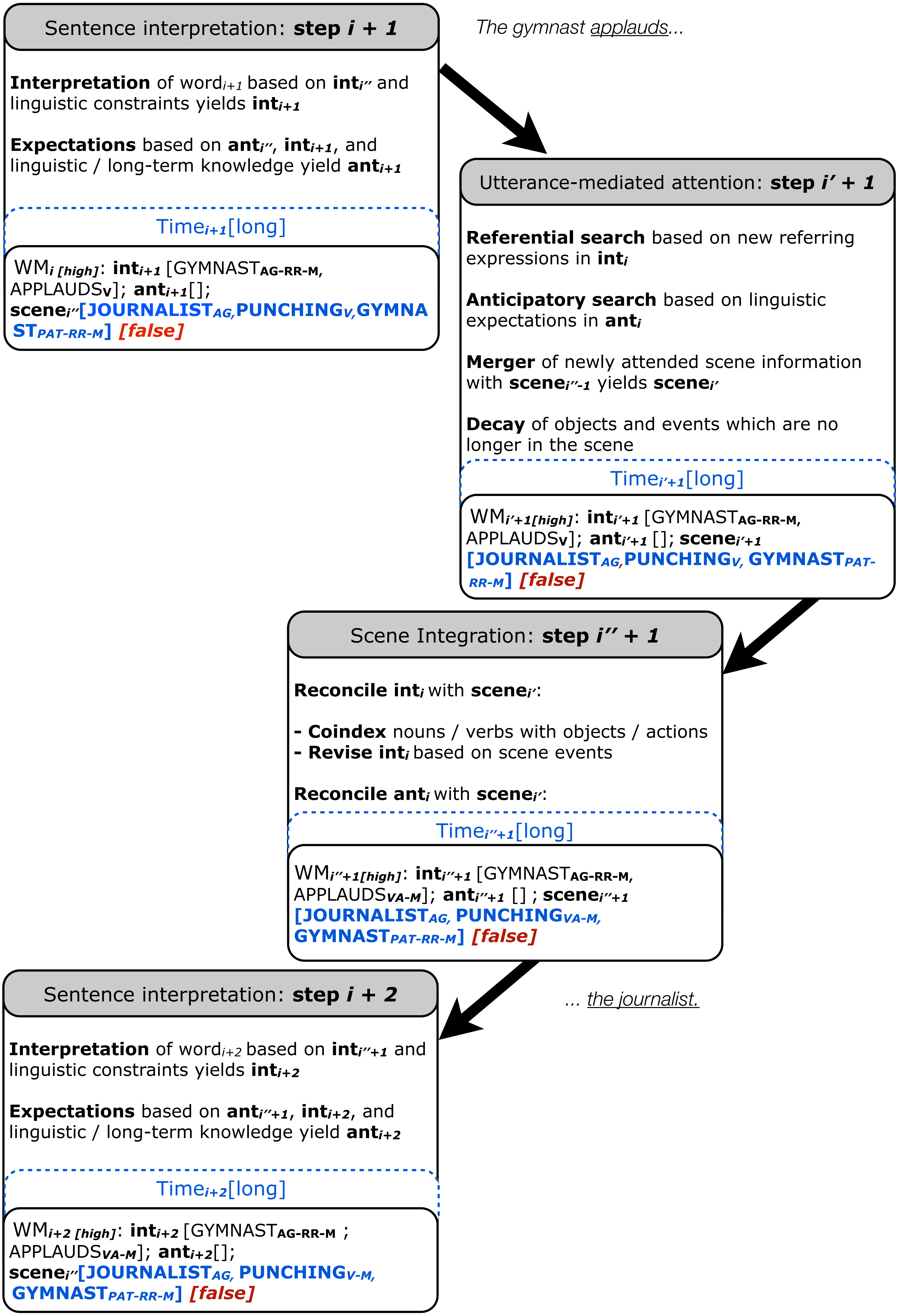
Interpretation of word$_{i+1}$ based on **int**$_{i''}$ and linguistic constraints yields **int**$_{i+1}$

Expectations based on **ant**$_{i''}$, **int**$_{i+1}$, and linguistic / long-term knowledge yield **ant**$_{i+1}$

Time$_{i+1}$[long]

WM$_{i\,[high]}$: **int**$_{i+1}$ [GYMNAST$_{AG\text{-}RR\text{-}M}$, APPLAUDS$_V$]; **ant**$_{i+1}$[]; **scene**$_{i''}$**[JOURNALIST$_{AG}$, PUNCHING$_V$, GYMNAST$_{PAT\text{-}RR\text{-}M}$]** ***[false]***

Figure 10



Sentence interpretation: **step *i* + 1**
The gymnast applauds...
Interpretation of word$_{i+1}$ based on **int**$_{i''}$ and linguistic constraints yields **int**$_{i+1}$
Expectations based on **ant**$_{i''}$, **int**$_{i+1}$, and linguistic / long-term knowledge yield **ant**$_{i+1}$
Time$_{i+1}$[long]
WM$_{i\ [high]}$: **int**$_{i+1}$ [GYMNAST$_{AG\text{-}RR\text{-}M}$, APPLAUDS$_V$]; **ant**$_{i+1}$[]; **scene**$_{i''}$[JOURNALIST$_{AG}$, PUNCHING$_V$, GYMNAST$_{PAT\text{-}RR\text{-}M}$] *[false]*
Utterance-mediated attention: **step *i′* + 1**
Referential search based on new referring expressions in **int**$_i$
Anticipatory search based on linguistic expectations in **ant**$_i$
Merger of newly attended scene information with **scene**$_{i''-1}$ yields **scene**$_{i'}$
Decay of objects and events which are no longer in the scene
Time$_{i'+1}$[long]
WM$_{i'+1[high]}$: **int**$_{i'+1}$ [GYMNAST$_{AG\text{-}RR\text{-}M}$, APPLAUDS$_V$]; **ant**$_{i'+1}$ []; **scene**$_{i'+1}$ [JOURNALIST$_{AG}$, PUNCHING$_V$, GYMNAST$_{PAT\text{-}RR\text{-}M}$] *[false]*
Scene Integration: **step *i″* + 1**
Reconcile int$_i$ with **scene**$_{i'}$:
- **Coindex** nouns / verbs with objects / actions
- **Revise int**$_i$ based on scene events
Reconcile ant$_i$ with **scene**$_{i'}$:
Time$_{i''+1}$[long]
WM$_{i''+1[high]}$: **int**$_{i''+1}$ [GYMNAST$_{AG\text{-}RR\text{-}M}$, APPLAUDS$_{VA\text{-}M}$]; **ant**$_{i''+1}$ [] ; **scene**$_{i''+1}$ [JOURNALIST$_{AG}$, PUNCHING$_{VA\text{-}M}$, GYMNAST$_{PAT\text{-}RR\text{-}M}$] *[false]*
Sentence interpretation: **step *i* + 2**
... the journalist.
Interpretation of word$_{i+2}$ based on **int**$_{i''+1}$ and linguistic constraints yields **int**$_{i+2}$
Expectations based on **ant**$_{i''+1}$, **int**$_{i+2}$, and linguistic / long-term knowledge yield **ant**$_{i+2}$
Time$_{i+2}$[long]
WM$_{i+2\ [high]}$: **int**$_{i+2}$ [GYMNAST$_{AG\text{-}RR\text{-}M}$; APPLAUDS$_{VA\text{-}M}$]; **ant**$_{i+2}$[]; **scene**$_{i''}$[JOURNALIST$_{AG}$, PUNCHING$_{V\text{-}M}$, GYMNAST$_{PAT\text{-}RR\text{-}M}$] *[false]*