Cognition 200 (2020) 104251

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses

Yujing Huang[a,*], Jesse Snedeker[b]

[a] Department of Linguistics, Harvard University, MA 02138, USA
[b] Department of Psychology, Harvard University, MA 02138, USA

ARTICLE INFO

Keywords:
Argument reactivation
Growth curve analysis
Unaccusativity
Visual world paradigm

ABSTRACT

Many syntactic theories posit a fundamental structural difference between intransitive verbs with agentive subjects (unergative verbs) and those with theme subjects (unaccusative verbs). This claim garners support from studies finding differences in the online comprehension of these verbs. The present experiments seek to replicate one such finding using the visual world paradigm (Koring, Mak, & Reuland, 2012). We control for several factors that were uncontrolled in previous studies. We find no differences in the processing of unergative and unaccusative sentences in logistic regressions and cluster analyses. However, in growth curve analyses, modeled closely on the original paper, we find differences between the verb conditions that *appear* to be statistically significant but are unstable across experiments. A resampling analysis reveals that the growth curve analyses are highly anticonservative, suggesting that the earlier finding was a false positive. We conclude that there is no strong evidence that unaccusatives are processed differently from unergatives. We suggest that growth curve analyses only be used with visual world paradigm data when the underlying assumptions of the analysis can be validated via resampling.

1. Introduction

How the meaning of a verb is mapped to its syntactic structure has been an important question in theoretical linguistics, as well as in language acquisition and psycholinguistics. Many linguists propose that there is a consistent mapping between thematic roles (e.g. agent and theme) and syntactic positions (i.e. internal and external arguments) (Baker, 1988). For a transitive verb such as *kick*, this mapping seems straightforward. The agent (kicker) maps onto the subject and the theme (kickee) maps onto the object. These mappings are so robust that by 21 months of age, children can already apply them to interpret novel verbs (Gertner, Fisher, & Eisengart, 2006). When it comes to intransitive verbs, however, this picture becomes more complicated.

Intransitive verbs differ in what kind of roles their subjects take. For example, the subject of the verb *scream* takes an agent role, while the subject of the verb *fall* takes a theme role. This can be shown by adding an *-er* morpheme to each verb in (1). In (1a) we can call the boy who screamed a *screamer* while in (1b) it is strange to call the boy who fell a *faller*. This is because the boy in (1a) is agentive and is interpreted as the initiator of the screaming event, while the boy in (1b) is interpreted as a theme which is affected by the falling event rather than initiating it.

(1) a. The boy **screamed**.
b. The boy$_i$ **fell** t_i.

To preserve a simple mapping rule between the thematic roles and syntactic positions, linguists have hypothesized that for the verbs with theme subjects (unaccusative verbs), the subject originates from the internal argument position and moves to the subject position in surface structure, leaving a trace in the object position. (A trace is a linguistic element that occupies a syntactic position but does not have a phonological form.) In contrast, the agentive arguments of the unergative verbs begin in the same structural position as the subjects of transitive verbs. This proposal, for a structural difference between the verb types, is called the Unaccusative Hypothesis (Burzio, 1981, 1986; Perlmutter, 1978). Several linguistic phenomena are believed to reflect this syntactic difference, such as: there-insertion (Burzio, 1986; Hoekstra & Mulder, 1990), the causative alternation (Burzio, 1986), impersonal passivization (Abraham, 1986), resultative constructions (Simpson,

* Corresponding author.
E-mail address: yujinghuang@fas.harvard.edu (Y. Huang).

https://doi.org/10.1016/j.cognition.2020.104251
Received 11 December 2017; Received in revised form 27 February 2020; Accepted 29 February 2020
Available online 13 May 2020
0010-0277/

1983), and auxiliary selection in Germanic and Romance languages (Rosen, 1984 for Italian, Haider & Rindler-Schjerve, 1987 for Italian and German, Zaenen, 1988 for Dutch, Legendre, 1989 for French, i.a.). Taking the resultative constructions as an example, it is claimed that the unaccusative verbs (2a) but not the unergative verbs (2b) can form a resultative construction without a reflexive before the adjective (2c).[1]

(2) a. The river froze solid.

b. *The boy laughed hoarse.

c. The boy laughed himself hoarse.

Many researchers have proposed that the syntactic distinction in (1) results in a difference in how unaccusative and unergative verbs are processed, either because of the need to reactivate the unaccusative subject at the object position or because unaccusative sentences are more structurally complex. This claim has been supported by findings using a variety of paradigms (e.g. Bever & Sanz, 1997; Burkhardt, Piñango, & Wong, 2003; Friedmann, Taranto, Shapiro, & Swinney, 2008; Koring, Mak, & Reuland, 2012; Momma, Slevc, & Phillips, 2018; i.a.). Among these studies, the most informative ones are Burkhardt et al. (2003), Friedmann et al. (2008) and Koring et al. (2012) which all use temporally sensitive methods and thus provide insight into when such a processing difference might occur.

These studies look for evidence that the subject of an unaccusative verb is reactivated shortly after hearing the verb. On a theory in which unaccusative verbs have a trace after the verb, this reactivation can be interpreted as retrieving the argument indexed with the trace. Two of these studies used cross-modal priming paradigms. Burkhardt et al. (2003) found reactivation of unaccusative subjects 650 ms after the offset of the verb, and a reactivation of unergative subjects 100 ms after verb offset. Friedmann et al. (2008) found subject reactivation 750 ms after verb offset in unaccusatives but no reactivation for unergatives at 750 ms or 0 ms after verb onset. Koring et al. (2012) used the visual world paradigm instead of cross-modal priming. They found a late reactivation (950 ms after verb offset) for unaccusative verbs and an early reactivation for unergative verbs, one that emerged shortly after verb onset. Koring et al. (2012) interpreted this pattern as follows: The subject of the sentence must always be reactivated after the verb to be integrated into the argument structure. Reactivation is fast for unergative verbs because they are structurally simpler, but it is slow for the more complex unaccusative verbs.

Finding a systematic delay in the reactivation of unaccusative subjects, as these studies seem to do, lends support to theories, like the Unaccusative Hypothesis, which propose that there is a fundamental difference in the structure of these verbs. There are, however, reasons to hesitate before we accept this conclusion. First, across these three studies there are differences in the time course of reactivation that are difficult to reconcile. If unergative subjects are activated shortly after the verb begins (Koring et al.) and remain active 100 ms after the verb ends (Burkhardt et al.) then why aren't they active at verb offset (Friedmann et al.)? In addition, each of these studies had potential or known confounds, described below, that might account for the pattern of findings. Finally, as we will see, the most relevant of these studies, Koring et al. (2012), used a statistical technique, growth curve modeling, which may not be well suited to visual world data.

Given the theoretical importance of this phenomenon, we set out to conduct a close, but not exact, replication of the Koring et al. (2012) study. We had two goals in doing this: 1) We wanted to assess whether this data pattern was stable and robust, even when all potential confounds were removed, in the hope that we could build upon these findings to explore syntactic processing in children and persons with developmental disorders. 2) We wanted to assess whether the growth curve models used in Koring et al. (2012) would produce consistent findings and whether those findings would be confirmed by other analyses that might arguably be better suited to the data structure. We discuss each of these goals below.

1.1. The rationale for a close, but not exact, replication

While our studies used the same basic design and procedure used by Koring et al. (2012), we changed the stimuli to remove some potential confounds that were present in the previous experiments. Unergative and unaccusative sentences necessarily differ in their verbs. In the three prior studies, these two groups of sentences also differed in the subject nouns that were used and thus in the primed pictures or probe words (Burkhardt et al., 2003; Friedmann et al., 2008; Koring et al., 2012). The authors of these studies were aware of this problem and matched their stimuli on several relevant features. However, in each study there was at least one factor that is known to play a role in language processing but was not matched, opening the door to an alternate interpretation of their findings. For example, unergative verbs tend to be more imageable than unaccusative verbs. To explore the role that this might have played in these experiments, we asked participants to rate the imagability of the verbs in the published stimulus sets for Friedmann et al. (2008) and Koring et al. (2012). We found that in both cases the unergatives were higher in imageability than the unaccusatives (Friedmann et al. $N = 28$, unaccusatives = 4.07, unergatives = 5.62, $p < 0.01$; Koring et al. N = 28, unaccusatives = 4.44, unergatives = 5.96, $p < 0.01$. Data in Appendix 2). This difference might account for the prior findings: More imageable words are recognized more quickly (see Paivio, 1991 for review) and thus we should expect faster processing for the unergative verbs in these experiments, based on imageabilty alone. This in turn could result in the more rapid reactivation of the subject. Other factors that were unmatched in these studies included: the codeability of the target picture, the imageability of the probe word that indexed subject reactivation, or the complexity of the critical sentence after the verb.

In our study, we addressed potential confounds in two ways: First, we eliminated confounds related to the subject noun, the sentence continuation, and the pictures by using the same pictures and sentence frames across verb classes (counterbalanced between subjects). Confounds linked to the verb could not be eliminated in this way, since the verbs in the two classes must be different, so we addressed these issues by matching the two sets of verbs for imageability, frequency, and for their congruency with the subject noun.

We considered these changes in the original design to be minor improvements– if we had replicated the Koring et al. data pattern, these changes would increase our confidence that verb class was really the relevant variable. However, as we will see, we did not replicate this pattern (Section 2). For this reason, we conducted two additional experiments, one seeking to increase the sensitivity of our measure (Section 3) and another that changed our instructions to more closely match those in Koring et al. (2012) (Section 4).

All three experiments are close replications according to the standards in LeBel, Berker, Campbell, and Loving (2017). They are parallel to the original experiment in the conceptualization and implementation of the independent variable, in the general procedure, and in how the dependent variable is measured (see Appendix 6). While our stimuli were modeled on Koring et al. (2012), they were different both because they were in English and because we controlled for a greater number of variables. Thus, these experiments were not exact replications. This was consistent with our goals. When the function of the replication is to spot false positives due to sampling error, more exact replications are preferred. But when the function of the replication is to rule out artifacts

[1] There are verbs for which the syntactic patterns deviate from semantic intuitions. For example, verbs like *sparkle* have theme subjects but pattern with unergative verbs in auxiliary selection in Dutch and German. For those who are sympathetic to the Unaccusative Hypothesis this is seen as evidence that the underlying syntactic difference cannot be reduced to a conceptual one. For those who are skeptical of the hypothesis, these verbs undermine the claim that there is any simple mapping between thematic roles and syntax and thus provide evidence against the hypothesis.

and determine whether the finding generalizes as predicted, then close but inexact replications are better (Schmidt, 2009). Critically, the changes we made in the stimuli and population would not be predicted to alter the reactivation effect given the linking hypothesis presented by Koring et al. (2012). The differences in reactivation, on Koring et al.'s hypothesis, should be consistent within a verb class and should be present across languages, including English (Burkhardt et al., 2003; Friedmann et al., 2008).

1.2. Concerns about growth curve models for visual world data

Our second goal was to explore the stability of the growth curve analyses that were used by Koring et al. (2012). Growth curve analyses model changes over time. Thus, they seem like a promising method for linking our cognitive theories to the rich data provided by the visual world paradigm (Mirman, Dixon, & Magnuson, 2008). Perhaps for this reason, the use of growth curve analyses in visual world studies has spread quickly (see Mirman & Magnuson, 2009; Kukona, Fang, Aicher, Chen, & Magnuson, 2011; Brown, Salverda, Dilley, & Tanenhaus, 2011; Lee, Middleton, Mirman, Kalénine, & Buxbaum, 2013; Hadar, Skrzypek, Wingfield, & Ben-David, 2016; Pozzan, Gleitman, & Trueswell, 2016; Cane, Ferguson, & Apperly, 2017; i.a.). However, a close look at the analysis in Koring et al. (2012) reveals some potential problems.

To conduct a growth curve analyses for a visual world study, the data for every trial must be aligned at a particular time point in the sentence (e.g., verb offset). Time is measured in small windows relative to that synchronization point (ranging from 17 ms to 200 ms). The measure of interest is whether the participant is looking at a particular object (the matching picture) during each of these time windows. Trials are averaged together within a participant to get a proportion of looks to a given item during that window. A multilevel linear regression model is constructed which predicts change over time in this value for each participant (at level 1) and predicts these parameters across participants (at level 2). This model is typically constructed by adding higher level polynomials one-by-one to capture increasingly complex patterns of change over time. The effects of an independent variable (such as verb class) can be assessed both by adding a main effect of the variable to the model and by adding interactions between this variable and the time parameters. Koring et al.'s (2012) analysis followed this basic pattern but had a few unique features (e.g., two distinct but overlapping time windows were used and the dependent variable was a difference score in looks to the target between the match and nonmatch trials). We had four concerns about these models.

First, the Koring et al. (2012) models included multiple parameters to capture differences between the verb classes (three in the first time window and five in the second). Since no correction is made for multiple comparisons, the probability of getting a false positive in one such parameter is quite high. In the absence of a clear linking hypothesis between these time parameters and our theory of processing, there is a temptation to interpret any significant parameter as evidence for reactivation, and to keep adding parameters until such a difference emerges.

Second, like most growth curve models the Koring et al. (2012) analysis uses a linear linking function, and thus it incorporates the assumption that the error is normally distributed on a linear scale. This idealization is false in the limit for visual world paradigm data because the behavior itself is binary (i.e. at a given time point, the participant either looks at the target picture or not). As Jaeger (2008) noted treating categorical data as linear can cause spurious effects.

Third, their analysis, like other growth curve analyses of visual world paradigm data, collapses across items in the same condition rather than treating item as a random effect (see Clark, 1973 for discussion). Thus, these analyses cannot support generalizations about a population of items, which is precisely the kind of generalization that underlies the unaccusative hypothesis.

Finally, the model used in Koring et al. (2012) did not account for autocorrelation in the data. Growth curve analyses, like most statistical analyses, assume that errors across different data points are not correlated. Visual world paradigm data, however, is known to have strong correlations between adjacent time points–at any given moment you are very likely to be looking at the same thing that you were looking at 16–50 milliseconds earlier (Cho, Brown-Schmidt, & Lee, 2018). Failing to correct for this can produce spurious results.

These issues are not unique to the Koring et al. (2012) study. All four of these features were present in the paper that introduced growth curve models to the psycholinguistic community (Mirman, Dixon et al., 2008). Three of the features—the use of a linear linking function, averaging across items, and no modeling of autocorrelation—are present in most of the subsequent visual world studies that have used growth curve models. Thus, looking closely at these particular models may help us understand the advantages and perils of growth curve modeling for visual world data more generally.

In the studies that follow, we conducted standard growth curve analyses that closely mirrored those of Koring et al. (2012). Our goal in doing this was to assess the stability of the method and analysis by conducting a close replication. However, we also conducted two additional analyses which avoided these four problems: one was a logistic mixed model on large time windows, while the second was a cluster analysis to compare conditions across time. To preview our results, we found effects that appeared to be statistically significant in each of our growth curve analyses, but these effects were not the same as those in Koring et al. (2012), and they were not the same effects across our three studies (Sections 2, 3 and 4). We found no significant effects in the logistic mixed models or the cluster analysis in any of our experiments. This pattern led us to explore the validity of the growth curve analyses using a resampling analysis, or Monte Carlo simulation (Section 5).

2. Experiment 1

This study was modeled on Koring et al. (2012). Participants viewed a visual display while listening to an auditory sentence. Previous research has shown that our eye movements are affected by the conceptual content of language (Altmann & Kamide, 2004; Yee & Sedivy, 2006 i.a.). If we hear the word “geographer” and there is a picture on the screen that is associated with a geographer, such as a map, we will tend to look more at that picture than we would otherwise. Koring et al. (2012) found that looks to semantically associated pictures also occurred when an argument was re-activated. Specifically, shortly after participants heard the verb, they were more likely to look to the picture related to the subject noun. As we noted above, they found subtle differences in the timing of these looks which they suggested reflected differences in the syntactic complexity of the verbs: 1) immediately after the verb there was a larger quadratic component for unergatives, interpreted as earlier integration and 2) later there was a difference in the quartic component (4th power) which they interpret as later integration of the unaccusatives due to complexity.

2.1. Method

2.1.1. Subjects

Forty monolingual native English speakers from the Harvard community participated in the study and were given either course credit or a $5 payment. They all reported that they had normal or corrected to normal vision and normal hearing. All studies were approved by the Committee on the Use of Human Subjects (CUHS) at Harvard University, and informed consent was obtained prior to the participants' involvement in the research.

2.1.2. Materials

2.1.2.1. Selection of verbs. All unaccusative verbs in this study were non-alternating verbs. We excluded verbs which can be both unaccusative and causative (for example, *break* in 3) for two reasons:

(1) There is theoretical controversy about whether alternating and non-alternating unaccusative verbs involve the same syntactic mechanism (Chierchia, 2004); (2) Previous research suggests that the alternating category may behave differently from non-alternative unaccusatives in online processing (Friedmann et al., 2008).

(3) a. The vase broke.

b. The boy broke the vase.

Thirteen of the 20 unaccusative verbs and 15 of the 20 unergative verbs from our study were taken from previous processing studies (Agnew, van de Koot, McGettigan, & Scott, 2014; Friedmann et al., 2008; Koring et al., 2012). The rest of the verbs were classified in the same subcategory as one of the previously used verbs in VerbNet 3.2 (Kipper, Korhonen, Ryant, & Palmer, 2008) or Levin (1993). The new unaccusative verbs in our study met the criteria for unaccusative verbs in Friedmann et al. (2008), i.e. (1) ability to be used in there-insertion construction, (2) ungrammaticality with a direct object, and (3) inability to undergo passivization. The new unergative verbs were all unambiguously intransitive and meet the criteria for unergative verbs in Friedmann et al. (2008), i.e. (1) ungrammaticality in the there-insertion construction, (2) ungrammaticality in the resultative construction, and (3) inability to occur with a reflexive pronoun unless the reflexive pronoun is followed by a resultative.

The lemma frequencies of unaccusative verbs and unergative verbs were calculated from the Corpus of Contemporary American English (COCA; Davies, 2008). The log transformed mean frequencies of unaccusative and unergative verbs (3.53 and 4.15 respectively) did not differ significantly ($t = -1.45$, $p = 0.16$). The imageability of the verbs was determined by an Amazon Mechanical Turk (AMT) norming study (following the design of Paivio, Yuille, & Madigan, 1968). The mean imageability of unaccusative and unergative verbs (4.13 and 3.84 respectively) did not differ significantly ($t = -1.19$, $p = 0.24$).

2.1.2.2. Sentences and images. We paired our unaccusative and unergative verbs and placed them in the same sentence frame so that in each pair, the only difference between the sentences was the verb (see Table 1). For each pair of verbs there were two different frames and two different subject nouns so that the same pictures could be used as semantically related and unrelated items (for the full list of stimuli, see Appendix 1). By pairing our stimuli in this way, we were able to use the same sentences, subjects, and pictures across the four conditions thus removing potential confounds between the unaccusative and unergative conditions.

All of our test and control sentences had the same structure:

Cookie monster said that	the subject	prepositional phrase modifying the subject
Cookie monster said that	*the geographer*	*with a loud voice and quick temper*
adverb	verb (+a prepositional phrase)	a temporal clause
suddenly	*fell*	*when the boat lurched violently because of the storm*.

Our sentence frames were closely modeled on Koring et al. (2012). Each subject noun was followed by a modifier to ensure that the initial activation of the subject would decay before the verb. There was an adverb before each verb, a feature that is present in all but one of Koring et al.'s (2012) stimulus sentences. This design choice could be critical for observing early reactivation: when the participants hear the adverb, they may be able to infer that the verb is coming up and this could lead them to begin activating the argument in preparation for integration. The modifiers and the adverbs together were 11–20 syllables long (mean = 15.12). Neither the modifier nor the adverb was semantically related to the critical argument, the target image, the verb or the distractors. Following Koring et al. (2012), we included a temporal clause after the verb phrase to ensure that there was time to detect late reactivation of the subject noun and to distinguish it from sentence wrap up effects. We added a prepositional phrase after the verbs in those sentences where it seemed necessary based on argument structure of the verbs. The temporal clauses and the post-verbal prepositional phrases together were 10–19 syllables long (mean = 13.3). Neither of the temporal clause nor the prepositional phrase was semantically associated with the subject noun, the target image or the distractors. We also made sure that the verbs were not semantically related to either the subject noun or the target image. All the subject nouns were terms for occupations and hence animate.

In each trial, there were four images on the screen, all black and white line drawings from Bank of Standardized Stimuli (Brodeur, Guérard, & Bouras, 2014), Bonin, Peereman, Malardier, Méot, and Chalard (2003) and Szekely et al. (2004). We conducted an AMT norming study to control for the relatedness between the critical images and the subject noun. The mean relatedness for the related image was 4.50 on a scale from 0 to 5 and the relatedness for the paired non-related image was 0.41.

In the match condition, the picture that was semantically related to the subject noun appeared with three distractors. For example, on match trials when participants heard the sentence with *geographer*, they saw a map as in Fig. 1. In the mismatch condition (the control condition), this related image was replaced with one that was unrelated to the noun (e.g., the baton in Fig. 1). Critically, the picture that was the unrelated image for one set of participants was the related image for another group of participants who heard the same verb in a different frame with a different subject noun (e.g., conductor, see Table 1). Thus, half of the time the right panel of Fig. 1 was used in the match condition and the left in the mismatch, and half the time it was reversed.

With this design, we ensured that the sentence frames and the visual stimuli were exactly the same across the unaccusative and the unergative conditions and across the test and control conditions. Thus, any differences we might find between unaccusative and unergative conditions, would be due to the differences between these two verb categories rather than uncontrolled differences between the pictures, arguments or sentence frames.

We normed the average naturalness of the sentences on AMT and found no significant difference (Mann-Whitney-Wilcoxon Test, W = 3380.5, $p = 0.54$) between sentences with unaccusative verbs (Mdn = 5.28) and sentences with unergative verbs (Mdn = 5.28) on a 7-point scale. We also controlled the plausibility of the verb given the subject noun for unaccusative condition (Mdn = 5.83) and unergative condition (Mdn = 6.06) (Mann-Whitney-Wilcoxon Test, W = 638, $p = 0.12$).

We created four lists so that each participant only heard the same

Table 1
Illustration of sample stimulus.

Verb	Sentence	Match	Control
Unaccusative	Cookie Monster said that the **geographer** with a loud voice and quick temper suddenly ***fell*** when the boat lurched violently because of the storm.	Map	Baton
Unergative	Cookie Monster said that the **geographer** with a loud voice and quick temper suddenly ***screamed*** when the boat lurched violently because of the storm.	Map	Baton
Unaccusative	Cookie Monster said that the **conductor** with a loud voice and quick temper suddenly ***fell*** when the boat lurched violently because of the storm.	Baton	Map
Unergative	Cookie Monster said that the **conductor** with a loud voice and quick temper suddenly ***screamed*** when the boat lurched violently because of the storm.	Baton	Map

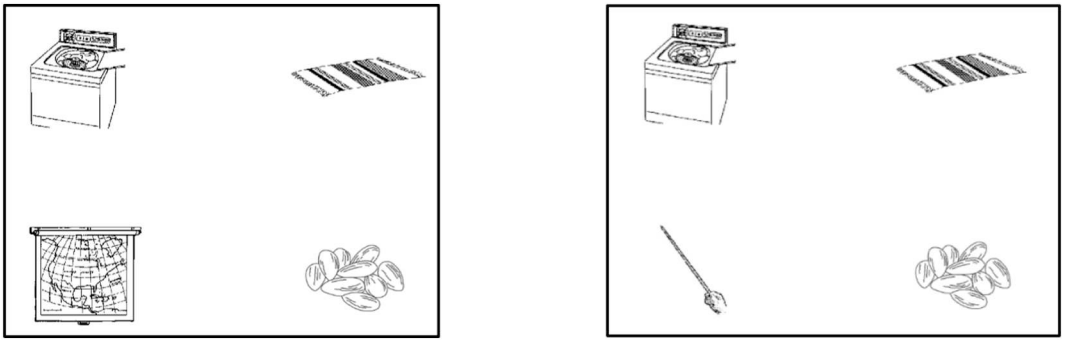

Fig. 1. Example visual displays for the sentences "Cookie Monster said that the geographer/conductor with a loud voice and quick temper suddenly fell when the boat lurched violently because of the storm." The targets are the *map* (left) and the *baton* (right).

sentence frame once, only encountered the same verb once, and only saw the same visual stimulus once. This was done by creating two sentence frames for each verb pair. In a given list, if sentence frame 1 appeared with an unergative verb, then the second sentence frame (for that pair) would appear with the unaccusative verb. This resulted in 40 critical sentences for each participant, 10 in each condition. There were 40 fillers which remained the same across all the four lists, resulting in a total of eighty sentences in each list. Our fillers, like those in Koring et al. (2012), had transitive verbs and were paired with displays in which one image depicted the subject of the sentence.

The audio stimuli were recorded at a normal speaking rate by a female native speaker of English, sampled at 44.1 kHz.

2.2. Procedure

Our procedure was closely based on Koring et al. (2012). Participants were seated comfortably in front of a monitor. Their eye movements were measured by a Tobii T60 sampling at 60 Hz. Each session started with a calibration procedure with seven fixation points. Participants were told that they would hear some sentences and look at some pictures. They were told to listen to the sentences carefully in order to answer some questions at the end of the study. Each trial started with a centrally-located fixation dot. Participants were instructed to look at the dot briefly. There was a one second preview of the display before the onset of the spoken sentence. After the sentence, there was a second of silence before the fixation dot appeared again. The entire experiment lasted about 20 min. At the end of the study, participants were given a memory test: they read to 32 sentences and were asked which ones they had heard in the study.

2.3. Results

2.3.1. The analytic strategy

First, we analyzed the data with a growth curve analysis that closely paralleled that of Koring et al. (2012). Second, we constructed simple cross-classified mixed effect models to predict the proportion of trials where participants primarily looked at the critical picture (the binarized mean fixation proportion, for more examples, see Hartshorne, Nappa, & Snedeker, 2015; Reuter, Feiman, & Snedeker, 2018; i.a.). This approach follows the rationale of the area-under-the-curve analysis which is widely used in analyzing visual world paradigm data (e.g. Chambers, Tanenhaus, & Magnuson, 2004; Papafragou, Hulbert, & Trueswell, 2008; i.a.), however it models looking time to a given target in a given time window on a given trial as binary, better capturing the distribution for short time windows. Doing these two analyses allowed us to compare the stability and informativeness of the models as well as the consistency of the results between the growth curve analysis and the traditional regression model. To preview our results: with the growth curve analysis, several terms were significant in our three replications of Koring et al. (2012), however, the specific terms and the directions of the effects were different across the studies. In contrast, in the logistic regression analyses, the three replications all gave negative results, meaning that we did not find a difference in looks to the semantic associate between the unaccusative and unergative conditions. Finally, we conducted cluster analyses to determine whether there were any smaller time windows in which the looking pattern significantly diverged for the two verb classes. This analysis allowed us to search for transient effects that might be lost by using large time windows. We found no evidence in the cluster analyses for differences between the verb classes. The cluster analysis was post hoc for Experiments 1 and 2, but planned for Experiment 3.

For all three studies presented in this paper, we preprocessed our data by removing trials with high track-loss. Specifically, we first excluded all samples with poor validity codes for both eyes (codes of 3 or 4) indicating that the tracker was unable to find either eye (as in the case of a blink) or had low confidence that it had done so. Only 19%–23% of the samples were lost in this way. Then we excluded any trial on which more than half of the expected samples were eliminated due to track loss. This resulted in the removal of between 1.3% and 1.5% of trials across the three experiments and two time windows. All the analyses were conducted in R (R Core Team, 2017) with the lme4 package (Bates, Maechler, Bolker, & Walker, 2015).

2.3.2. Growth curve models

Following the data processing procedure in Koring et al. (2012), we first calculated the proportion of looks to the target picture in each condition by participant in each 20 ms time bin (aggregated across items). Then, for each verb type, we subtracted the proportion of looks to the target picture in the mismatch condition from the proportion of looks to the target picture in the match condition by participant in each time bin. This difference score was used as the dependent variable. In a growth curve analysis, the change of the proportion of looks to the target over time is modeled by orthogonal power polynomials. Random slopes (by subject) were included for all time polynomials that were entered as fixed effects. We built up the model step by step: first we added the time polynomials, then the condition effect and then the interaction of condition with each of the time polynomials (see Table 2). Because these analyses are conducted on difference scores, an effect of *condition* (or interaction with condition) would indicate an underlying interaction of verb type and match.[2]

We analyzed our data using the two time-windows defined by Koring et al. (2012). The first window, the verb frame, was centered on

[2] Our model comparison procedure was different from Koring et al. (2012), where they built the model by adding both the higher order time polynomial term and the interaction in the same step. We made this choice to ensure that our model comparisons were relevant to the critical hypothesis (is there a difference between the verb classes in how the effect emerges over time). Nevertheless, the final models we report have the same variables as those of Koring et al. (2012) and thus are directly comparable.

Table 2
Model comparison of verb window for all experiments. The asterisks are used to flag levels of significance: p < 0.05 (*), p- < 0.01 (**), and p < 0.001(***).

Model	Experiment 1			Experiment 2			Experiment 3		
	AIC	Chisq	p-Value	AIC	Chisq	p-Value	AIC	Chisq	p-Value
Linear	−1122.30			−2761.60			−1904.30		
linear + quadratic	−1253.70	139.36	< 0.001***	−3194.00	440.38	< 0.001***	−2118.10	221.84	< 0.001***
linear + quadratic + condition	−1254.00	2.28	0.1	−3193.70	1.74	0.19	−2117.30	1.17	0.28
linear + quadratic + condition + linear*condition	−1264.80	12.78	< 0.001***	−3192.30	0.56	0.45	−2125.80	10.55	< 0.01**
linear + quadratic + condition + linear*condition + quadratic*condition	−1266.40	3.60	0.06	−3209.60	19.26	< 0.001***	−2136.4	12.54	< 0.001***

the verb offset (plus 200 ms, to account for the time it takes to program an eye-movement). It started 600 ms before verb offset and ended 1000 ms after verb offset. The second window, the post-verb frame, started 200 ms after verb offset and ended 1700 ms after verb. Thus our windows, like those in Koring et al., overlap.

2.3.2.1. Verb frame. In this window, the difference score was modeled by condition (unaccusative vs. unergative), linear term, quadratic term and their interactions with condition. The unergative condition was coded as the baseline. The results of the model comparisons are summarized in Table 2 (on the left). Adding terms did not always improve the model. Critically, the model with a quadratic interaction did not fit better than the model with a linear interaction. Nevertheless, we report the results of full quadratic model because it is the model used in Koring et al. (2012).

The results from the full quadratic model are summarized in Table 3, along with the Koring et al. findings. The critical effect in the Koring et al. (2012) study was the interaction between the quadratic component and condition. In the unergative condition, they found a significant negative quadratic component—a rise followed by a fall—-which they interpreted as evidence for early reactivation. Their unaccusative condition had a positive quadratic component–a fall followed by a rise. In contrast, in our model the interaction between condition and the quadratic term was not significant. ($p = 0.06$). But even more critically, the coefficient for this effect was negative, given our coding scheme this means that it was in the opposite direction of the effect found in the Koring et al. analysis (more rise-fall in the unaccusative condition than in the unergative, see Fig. 2). The models diverge in other parameters as well: Koring et al. find a robust main effect of condition, we do not; we find a large interaction between condition and the linear term, while they do not.

2.3.2.2. Post-verb frame. In this window, the difference score was modeled by condition, linear term, quadratic term, cubic term, quartic term and their interactions with condition. As can be seen in Table 4, adding the condition and its interaction with the higher order time terms did not consistently improve the model fit. Nevertheless, to facilitate comparisons with Koring et al., we report the results for the full quartic model in Table 5. Again, our findings diverge from Koring et al.'s (2012) in critical ways. Both studies find a significant interaction between the linear component and the condition, but the two effects go in opposite directions. Unlike Koring et al. (2012), we found a significant interaction of quadratic term and condition, but no interaction between the quartic term and condition.

2.3.3. Logistic mixed effect model

This analysis used the same two time windows as the growth curve models. Within each window for each trial, we calculated the average proportion of looking time to the target picture. Because saccades typically occur only once or twice per second (and many are within the same quadrant), on most trials the proportion of target looking time was either 1 or 0. Thus a binomial model was most appropriate. For this reason, we processed our data to completely binarize our variable: if the participant looked at target > 30% of the time (on that trial, in that time window) we coded it as 1 (a target look) otherwise we coded it as 0. Below we call this dependent variable "Target Looks".

Fig. 4 shows the proportions of trials with target looks in the verb and post-verb windows. In both regions, Target Looks were higher in the match condition (e.g. map given the subject geographer) than the mismatch condition (baton given geographer). Our central question was whether the size of the match effect is larger for one verb class than the other (suggesting greater reactivation). To test this, we constructed a logistic mixed effect model for each time window with Target Looks as the dependent variable and verb type (unaccusative = 1, unergative = −1), match condition (match = 1, mismatch = −1) and their interaction as fixed effects. Participants and items where included

Table 3
Growth Curve Analysis results in the verb frame for all experiments in comparison to Koring et al. (2012). The entries in green (boldface) have significant negative coefficients, while the entries in red (italics) have significant positive coefficients. The asterisks are used to flag levels of significance: p < 0.05 (*), p < 0.01 (**), and p < 0.001(***).

	Koring et al.			Experiment 1			Experiment 2			Experiment 3		
Parameter	**β**	**t**	**p-value**	**β**	**t**	**p-value**	**β**	**t**	**p-value**	**β**	**t**	**p-value**
Intercept*Condition	**-0.08**	**-2.21**	**<.05***	0.008	1.51	.13	-0.005	-1.32	<.19	0.006	1.11	<.27
Linear*Condition	-0.02	-0.1	n.s.	**-0.17**	**-3.58**	**<.001*****	0.03	0.74	.46	**-0.15**	**-3.27**	**<.01****
Quadratic*Condition	*0.20*	*7.89*	*<.001****	-0.09	-1.90	.06	**-0.16**	**-4.39**	**<.001*****	*0.16*	*3.54*	*<.001****

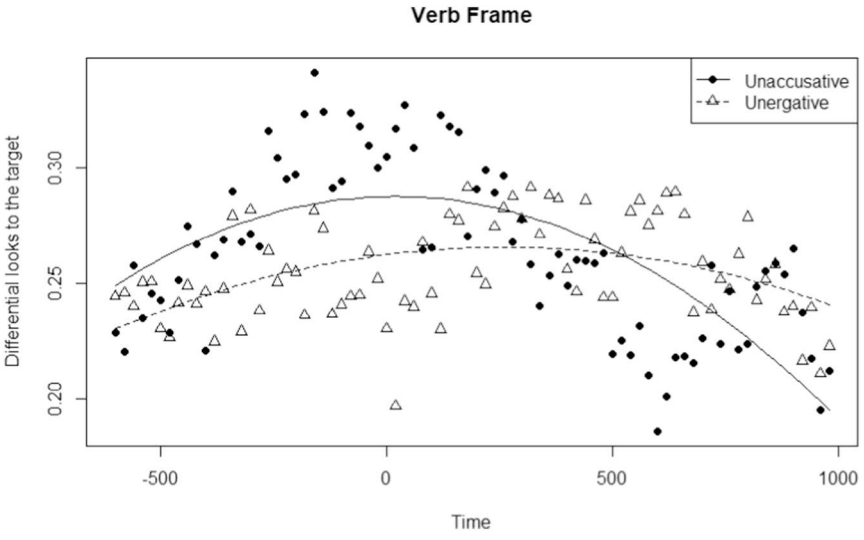

Fig. 2. Experiment 1: Fitted lines for the differential looks to the target in the two conditions. Differential looks is the difference between the proportion of looks to the target in the match condition and the proportion of looks to the target in the mismatch condition. Zero is the verb offset. The individual points correspond to the average of the dependent variable (across subjects and items) at each time point (every 20 ms).

as random intercepts with random slopes by participant and item for match condition and verb type.[3] Model comparison showed that this random effect structure had better fit than the simpler models with no random slopes in both windows (p's < 0.001). Models with more complicated random effects structure (e.g., random slopes for the interaction term) did not converge. The results are summarized in Table 6.

For the verb window, there was a significant effect of match condition ($p < 0.001$): there were more trials with looks to the target picture in the match condition than in the mismatch condition. There was no effect of verb type ($p = 0.79$), and critically there was no interaction between verb type and match ($p = 0.66$). Thus, we found no evidence for a difference in the activation of the subject across the two verb types.

The same model was run in the post-verb region. There was a significant effect of match (p < 0.001) and of verb type (p < 0.05). But critically there was no interaction between the two variables ($p = 0.81$).

In summary, we found no evidence for the critical interaction in either time window. If participants re-activated the subject of the unergative verbs more quickly, then, in the verb frame, we might have expected to see more looks to the match in the unergative condition than in the unaccusative condition, resulting in an interaction (with a negative beta). Similarly, if participants showed later reactivation of the unaccusative verbs, we might have expected to see a positive interaction in the post-verb frame. Neither effect was present. However, it is possible that these analyses were too coarse to detect fleeting differences in the timing of reactivation. For example, visual inspection of Figs. 2 & 3 could lead one to wonder whether the critical interaction is present in smaller time windows (e.g., around 500–700 ms after the verb). We address this possibility in our third analysis.

2.3.4. Cluster analysis

There was a clear discrepancy between the result of the growth curve analysis and that of the logistic regression. The growth curve analysis suggested that there were differences between the unaccusative and unergative conditions in both analysis windows, but the logistic regression did not. One explanation for this discrepancy is that the growth curve analysis produced spurious effects due to the limitations described in the introduction. A second possibility is that the divergence is due to differences in the temporal resolution of the two analyses. The logistic regression collapses the data across a long time window and thus might not capture fleeting effects, while the growth curve analysis retains information about the timing of eye-movements which may allow it to detect short-lived effects. To tease apart these possibilities, we conducted a third analysis which has a finer temporal resolution but avoids the problematic assumptions of growth curve analyses. Specifically, we conducted a permutation cluster analysis modeled on prior work in EEG (see Maris & Oostenveld, 2007) and in the visual world paradigm (Hahn, Snedeker, & Rabagliati, 2015).

For the cluster analysis, we first conducted a separate test for the critical interaction at each individual time point, we then linked together clusters of adjacent time points where there was an effect, and

[3] We considered each frame to be an item and consequently both verb type and match condition were manipulated within items.

Table 4
Model comparisons for post-verb window for all experiments. The asterisks are used to flag levels of significance: p < 0.05 (*), p- < 0.01 (**), and p < 0.001(***).

Model	Experiment 1			Experiment 2			Experiment 3		
	AIC	Chisq	p-Value	AIC	Chisq	p-value	AIC	Chisq	p-Value
linear + quadratic + cubic	−760.70			−2026.10			−2178.90		
linear + quadratic + cubic + quartic	−829.40	80.64	< 0.001***	−2168.60	154.55	< 0.001***	−2250.60	83.63	< 0.001***
linear + quadratic + cubic + quartic + condition	−830.57	3.21	0.07	−2171.50	4.92	< 0.05*	−2248.80	0.28	0.59
linear + quadratic + cubic + quartic + condition + linear*condition	−833.04	4.47	< 0.05*	−2170.50	0.96	0.33	−2253.60	6.73	< 0.01**
linear + quadratic + cubic + quartic + condition + linear*condition + quadratic*condition	−837.43	3.6.38	< 0.05*	−2185.30	16.82	< 0.001***	−2253.20	1.65	0.20
linear + quadratic + cubic + quartic + condition + linear*condition + quadratic*condition + cubic*condition	−838.31	2.89	0.09	−2186.20	2.87	0.1	−2256.10	4.92	< 0.05*
linear + quadratic + cubic + quartic + condition + linear*condition + quadratic*condition + cubic*condition + quartic*condition	−836.41	0.09	0.76	−2184.60	0.40	0.53	−2254.70	0.57	0.45

finally we corrected for multiple comparisons by conducting a non-parametric permutation test to determine the *p*-value for a given cluster size. The rationale is that, if there is a series of consecutive time bins that show a significant interaction between verb type and match, and if the number of the consecutive time bins is larger than what we would observe in a null distribution, then we can be fairly confident that the match effect is different for the two verb types during that time window.

This analysis has one critical advantage over both of our earlier analyses: it is less sensitive to how we define our window for analysis. In the logistic analysis, we collapse all data within a time window and thus we can fail to find a difference because we are collapsing a period in which an effect occurs with a period in which it is absent or reversed. In a growth curve analysis, the trajectory of the curve is highly dependent on our starting point and end point. This is particularly worrisome, in cases like this one, where there is no obvious criterion for defining the time window of interest. In contrast, a cluster based permutation analysis can find effects wherever they occur and considers each time point in isolation. Thus, it is best to start with a long window that captures the entire period over which the effect might be present, and then use the analysis to determine whether there is a smaller window in which the effect appears.

The analysis window we chose began 600 ms prior to the verb offset and ended 2000 ms after the verb offset. We chose the starting point because it corresponded to the beginning of the verb (which had an average duration of 600 ms) and thus represented the earliest point at which the sentences could possibly diverge. We chose to end our window for analysis 300 ms later than Koring et al.'s post-verb window to ensure that any clusters corresponding to the end of their analysis would not be artificially truncated. Critically, even the shortest of the stimulus sentences were still playing at 2000 ms.

Clusters were defined as the number of consecutive time bins of which the critical predictor had a *p*-value smaller than the predetermined threshold (0.05 in this case). To find the clusters, we first grouped our data into consecutive time bins of 100 ms and binarized our variable in each bin as described in Section 2.3.2 (using 30% as our threshold). In practice, the vast majority of values were 0 or 1 prior to binarization. Then, we ran the logistic mixed effect model described in Section 2.3.3 on each time bin. Because of the large number of analyses, and problems of convergence, this analysis included random intercepts for subjects and items but did not include random slopes. For each time step, we assessed whether there an effect of the critical interaction term that was significant at $p < .05$.

To figure out the likelihood of finding a cluster of a given size in a null distribution, we performed a permutation test. The goal of a permutation test is to retain as much of the co-variance structure of the data as possible while permuting the critical dependent variable to determine the empirical null distribution. For this reason, the unit that we permuted was the item. In our design, item referred to a sentence frame and a picture set, and each item could appear in one of four conditions (verb type × match). We expected that looking patterns would be strongly determined by the pictures themselves and the words in the utterance, independent of the nature of the verb or subject reactivation. Thus, randomizing the critical variables on an item-by-item basis would preserve this structure while testing the null hypothesis with respect to verb class.

Specifically, for each permutation we flipped the verb type label for a randomly selected set of half of the items (making the unergatives into unaccusative and vice versa). We then conducted our critical analysis on every time bin in the permuted sample (testing for verb type by match interactions). We grouped adjacent time points with significant effects into clusters and put these clusters aside. We performed the permutation 1000 times to create an empirical distribution of the number of clusters, of different sizes, in samples taken from the null distribution. This allowed us to determine the *p*-value for any clusters that we found in the actual data.

Table 5
Growth Curve Analysis results in the post-verb frame for all experiments in comparison to Koring et al. (2012). The entries in green (boldface) have significant negative coefficients, while the entries in red (italics) have significant positive coefficients.

	Koring et al (UE-UA)			Experiment 1			Experiment 2			Experiment 3		
Parameter	**β**	**t**	**p-value**	**β**	**t**	**p-value**	**β**	**t**	**p-value**	**β**	**t**	**p-value**
Intercept*Condition	-0.05	-1.32	n.s.	-0.01	-1.80	<.07	*0.01*	*2.21*	*<.05**	0.003	0.53	<.59
Linear*Condition	*0.42*	*2.73*	*<.01***	**-0.05**	**2.11**	**<.05***	-0.04	-0.97	.33	*0.11*	*2.60*	*<.01***
Quadratic*Condition	-0.03	-1.17	n.s.	**-0.09**	**2.53**	**<.05***	**-0.15**	**-4.10**	**<.001*****	0.06	1.28	0.20
Cubic*Condition	-0.01	-0.48	n.s.	0.08	-1.70	.09	0.06	1.69	.1	*0.10*	*2.22*	*<.05**
Quartic*Condition	*0.10*	*3.72*	*<.001****	0.17	0.31	.76	0.02	0.63	.53	0.03	0.76	0.45

To summarize, the steps of the cluster analysis are:

(1) For the critical predictor (the verb type by match interaction), find clusters of temporally-adjacent samples where the p-value is smaller than some predetermined threshold. The cluster size can be as small as one.
(2) Permute the data by randomizing the verb type label within each item while maintaining the data structure in all other respects.
(3) Run step (1) on the permuted data, extracting the number of clusters that are equal to or larger than the size of the cluster found in the original sample.
(4) Run steps (2) and (3) 1000 times to create an empirically determined null distribution.
(5) Take the data from the original sample and compare it to this null distribution. The likelihood of finding a cluster that is equal to or larger than the size found in the original data is calculated as the proportion of permuted samples that contains a cluster of the same size or larger in the 1000 samples.

Our initial plan was to conduct this analysis using an alpha of 0.05 in each time bin. However, there were no time points in our data (step 1) where the p-value of the verb-type by match interaction was < 0.05 and thus there could not be any cluster that was significant at the 0.05 level. This finding should impact how we interpret Figs. 2 and 3. In looking at these figures, one is tempted to think that there are short lived time windows where the two verb classes diverge, because there are stretches of time where the unergative points are all lower than the unaccusative points. This visual logic has some basis when interpreting a scatterplot, where each dot is an independent observation. In the present case, however, each point in the figure represent a time point. These points collapse across many different verbs, and many different trials, disguising the variability between them. But every point within a verb class, is composed from the same set of trials and thus tightly yoked to the one before it. The analysis above tells us that there is not a single time point, in either figure, where the unergative verbs are significantly different from the unaccusatives.

We conducted a secondary cluster analysis with an alpha of 0.2 in each time window to explore whether there might be a weak but long-lasting effect that the analysis above could not capture (Hahn et al., 2015). Because the final p-value of the cluster-based permutation test is determined not by the alpha used to evaluate each time point, but rather by the empirical distribution of clusters in the permutation test, doing this does not change the probability of a false positive. At this relaxed threshold, we found four consecutive time bins in which the interaction term had a p-value smaller than our criterion. This window began 200 ms before verb offset and ended 200 ms after the verb offset.

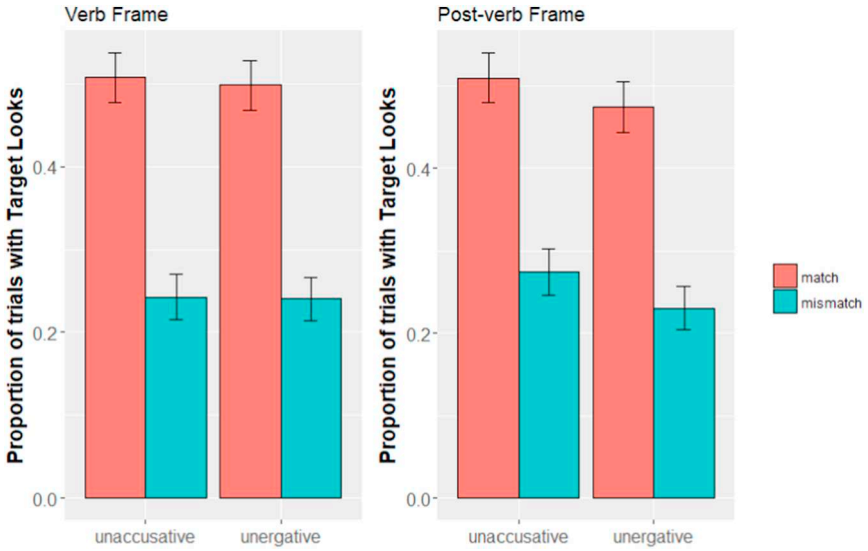

Fig. 4. Experiment1: Proportions of trials with target looks in the verb window (left) and post-verb window (right). The error bars reflect the 95% confidence intervals.

Table 6
Logistic regression results in the verb frame for all experiments. The asterisks are used to flag levels of significance: p < 0.05 (*), p- < 0.01 (**), and p < 0.001(***).

Parameter	Experiment 1				Experiment 2				Experiment 3			
	β	SE	z	p-Value	β	SE	z	p-Value	β	SE	z	p-Value
Intercept	−0.23	0.10	−2.21	< 0.05*	−0.60	0.07	−8.73	< 0.001***	−0.42	0.11	−3.95	< 0.001***
Match	0.59	0.09	6.36	< 0.001***	0.20	0.06	3.49	< 0.001***	0.47	0.04	11.54	< 0.001***
Verb type	−0.01	0.05	−0.27	0.79	−0.06	0.04	−1.55	0.12	−0.01	0.04	−0.31	0.76
Match*verb type	0.02	0.04	0.44	0.66	0.00	0.03	0.14	0.89	−0.03	0.04	0.72	0.47

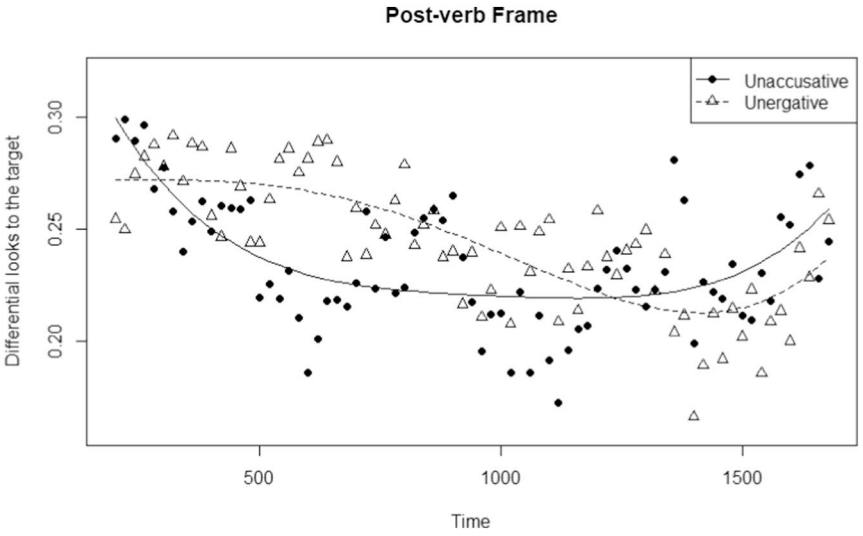

Fig. 3. Experiment 1: Fitted lines for the differential looks to the target in the two conditions. Differential looks is the difference between the proportion of looks to the target in the match condition and the proportion of looks to the target in the mismatch condition. Zero is the verb offset. The individual points correspond to the average of the dependent variable (across subjects and items) at each time point (every 20 ms).

This cluster, however, was no larger than what we might expect by chance from the null distribution: among the 1000 permuted samples, we found 467 clusters of size four or larger, given an alpha of 0.2, resulting in $p = 0.467$.

In sum, the cluster-based permutation analyses provide no reason to believe that there is any difference in the time course of subject reactivation for unaccusative and unergative verbs. These findings are consistent with the logistic regression but at odds with the results of the growth curve analysis.

2.3.5. Interim summary

In Experiment 1, we failed to find any evidence for a difference in subject reactivation for unergative and unaccusative verbs. We considered three reasons why our findings might diverge from those of Koring et al. (2012). First, the original finding could be a true positive and our failure to find the effect could be a fluke. To explore this possibility, we conducted a second study, with a larger sample, that was designed to get a larger reactivation effect, in hopes that we might find the expected data pattern (Experiment 2, Section 3). Second, the original finding could be a true positive and we may have failed to find it due to a small change that we made in the experimental procedure (telling participants that they would be asked questions at the end). We pursued this possibility in Experiment 3 (Section 4). Finally, the original finding could be false positive attributable to limitations of the growth curve analyses used in that paper. In Section 5, we explore this hypothesis by conducting resampling analyses to explore the rate of false positives in these analyses.

3. Experiment 2

In Experiment 2, we made a second attempt to reproduce the Koring et al. (2012) data pattern. We made two changes that we thought might increase our chances of detecting any effect. First, we added comprehension questions between some of the items to disguise the purpose of the study and encourage our participants to attend to the sentences. Second, we increased the number of subjects to 60 in hopes of increasing our power (Koring et al., tested 37).

3.1. Subjects

Sixty monolingual native English speakers were recruited from the Harvard community. They received either course credits or $5 payment as a compensation. They all reported that they had normal or corrected to normal vision and normal hearing.

3.2. Materials and procedure

The procedure and stimuli were the same as those of Experiment 1 except that 16 questions about the sentences or the pictures were interspersed among the test trials. For each question, two choices were provided. Participants needed to use the touch screen to select an answer. Once a choice was made, the study would proceed. This modification was made because in the debriefing procedure of Experiment 1, several participants mentioned that they thought the study was about the target picture and deliberately fixed their eyes on those pictures for a long time.

3.3. Results

3.3.1. Growth curve analysis

The analysis employed the same modeling procedure as in Experiment 1. Critically, once again, we forced the models to contain the same parameters that were used in Koring et al., (2012). These results appear in Tables 2-5 and Figs. 5 and 6.

In the verb window there was no interaction of condition and the linear time term ($p = 0.46$). There was a significant interaction of condition and the quadratic term ($p < 0.001$), but the sign for this interaction was negative (more rise and fall for unaccusatives, Fig. 5) while

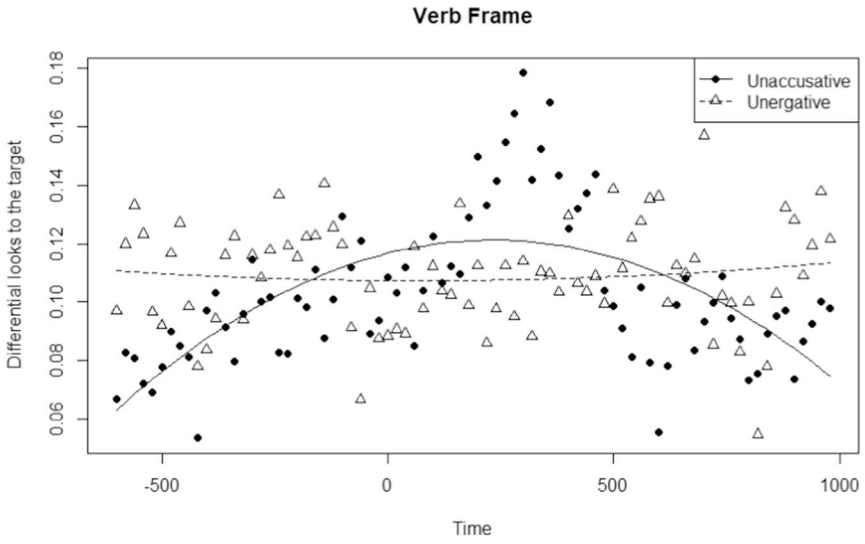

Fig. 5. Experiment 2: Fitted lines for the differential looks to the target in the two conditions. Differential looks is the difference between the proportion of looks to the target in the match condition and the proportion of looks to the target in the mismatch condition. Zero is the verb offset. The individual points correspond to the average of the dependent variable (across subjects and items) at each time point (every 20 ms).

in Koring et al. (2012) it was positive (see Table 3). In the post-verb window, there was a significant interaction of condition and the quadratic term ($p < 0.001$). All the other interactions between the condition and time terms were not significant (all p's $>$ 0.05). In contrast, Koring et al. (2012) found a significant interaction between the quartic term and the condition but no interaction between condition and the quadratic term.

Critically, the pattern of effects in Experiment 2 is also different from the pattern observed in Experiment 1. This can be seen most clearly in Tables 3 and 5. This suggests that the results of the growth curve analyses are highly unstable across closely parallel studies.

3.3.2. Logistic mixed effect model

We ran the same logistic regression models as in Experiment 1 (see Tables 6 and 7 and Fig. 7). In the verb window, the results closely paralleled Experiment 1: there was a significant match effect ($p < 0.001$), no verb type effect ($p = 0.12$) and no interaction between these two terms ($p = 0.89$). In the post-verb window, we again found a main effect of match ($p < 0.05$) but no interaction of verb type and match ($p = 0.60$). In Experiment 2, the main effect of verb type was not significant ($p = 0.38$).

3.3.3. Cluster analysis

As in Experiment 1, the results from the growth curve analysis and those of the logistic regression are at odds. To determine whether this may have been the result of the coarse temporal resolution of the logistic regression, we again conducted a cluster analysis, following the same procedure described above. For Experiment 2, there was (again) no data point with an interaction at an alpha of 0.05. With an alpha of 0.2 there were three consecutive time points at the end of the analysis window (1600–1900 ms after the verb offset). In the 1000 permuted samples, with an alpha of 0.2, there were 648 clusters of size three or larger and thus we can conclude that differences of this size are expected under the null hypothesis.

4. Experiment 3

In both Experiment 1 and 2, participants were told that they should listen to the sentences because they would be asked to answer some questions. This detail of our procedure differed from Koring et al. (2012) where there was no explicit task. It has been suggested that this change might account for the difference in findings between our studies and theirs. To address this possibility, we conducted Experiment 3 in which we eliminated this instruction (and the subsequent questions) to more precisely mirror the procedure of Koring et al. (2012).

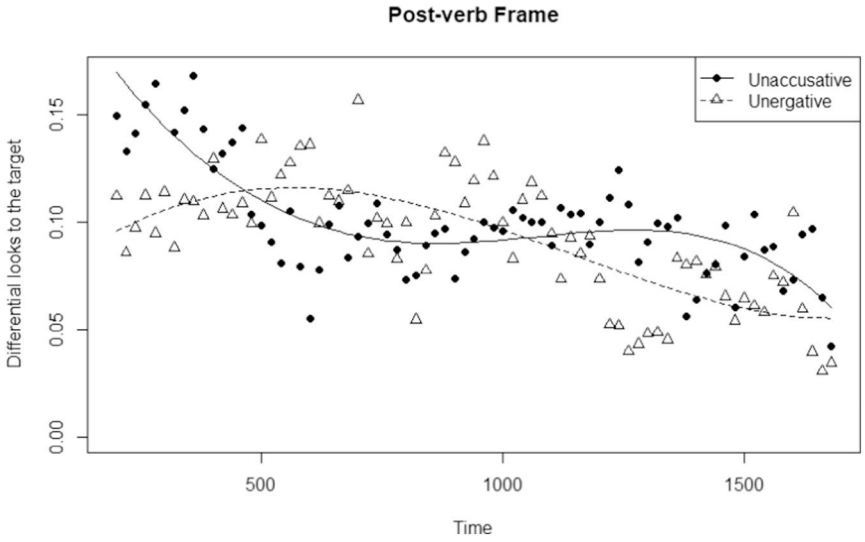

Fig. 6. Experiment 2: Fitted lines for the differential looks to the target in the two conditions. Differential looks is the difference between the proportion of looks to the target in the match condition and the proportion of looks to the target in the mismatch condition. Zero is the verb offset. The individual points correspond to the average of the dependent variable (across subjects and items) at each time point (every 20 ms).

Table 7
Logistic regression results in the post-verb frame for all experiments. The asterisks are used to flag levels of significance: p < 0.05 (*), p- < 0.01 (**), and p < 0.001(***).

Parameter	Experiment 1				Experiment 2				Experiment 3			
	β	SE	z	p-Value	β	SE	z	p-Value	β	SE	z	p-Value
Intercept	−0.25	0.10	−2.48	0.05	−0.63	0.07	−8.90	< 0.001***	−0.42	0.12	−3.57	< 0.001***
Match	0.52	0.09	5.62	< 0.001***	0.19	0.06	3.20	< 0.05*	0.45	0.09	4.81	< 0.001***
Verb type	0.11	0.05	2.09	< 0.05*	−0.04	0.04	−0.82	0.38	0.05	0.05	0.94	0.35
Match*Verb type	−0.01	0.04	−0.24	0.81	0.02	0.03	0.52	0.60	0.01	0.04	0.30	0.76

4.1. Subjects

Forty monolingual native English speakers were recruited from the Harvard community. They received either course credits or $5 payment as a compensation. They all reported that they had normal or corrected to normal vision and normal hearing.

4.2. Materials and procedure

The procedure and stimuli were the same as those of Experiment 1 and 2 except that all the questions were removed, and the instructions for the subjects were directly translated from Koring et al. (2012).

4.3. Results

4.3.1. Growth curve analysis

The analysis followed the same modeling procedure as we specified in Experiment 1 and 2. Once again, we forced the models to contain the same parameters that were used in Koring et al. (2012). These results appear in Tables 2–5 (final column).

In the verb window, there was a significant interaction of the quadratic term and condition (p < 0.001) (see Fig. 8 and Table 3) which was in the same direction as Koring et al. (2012). There was also a significant interaction of the linear term and condition ($p < 0.01$) which was not found in Koring et al. (2012). (See Fig. 9.)

In the post verb window, there was a significant interaction of linear term and condition which was also found in Koring et al. (2012). There was also a significant interaction of the cubic term and the condition which was not found in Koring et al. (2012). The interaction of the quartic term and condition that was significant in Koring et al. (2012) was not significant in Experiment 3.

Thus, in this close replication, using the Growth Curve analysis, we find a mix of effects, some which match those in the Koring et al. (2012) and some which do not. In the absence of a theory about the psychological significance of linear, quadratic, cubic and quartic effects, it is hard to know what one would make of such a pattern, if it were real. We will argue instead (Sections 5 & 6) that these apparent effects are consistent with the null hypothesis (that there are no differences between the two classes of verbs on these measures, in this task).

4.3.2. Logistic mixed effect model

We ran the same logistic regression models as in Experiment 1 and 2 (see Table 6 and Fig. 10). In both the verb window and the post verb window, there was a significant match effect (p's < 0.001) but no verb type effect (p's > 0.05) and no interaction between these two terms (p's > 0.05). Thus, the findings of this analysis are consistent across the three experiments.

4.3.3. Cluster analysis

We conducted a cluster analysis following the same steps specified in Experiment 1 and 2. There was no time point with a significant interaction between match and verb type at either the 0.05 level or 0.2 level. Thus, there was no potential effect to compare to the chance distribution. Critically, this analysis provides another illustration of how differences that are visually salient in the figures for the GCA can be meaningless. In Fig. 8, there is a cluster of points where the two verb classes appear to diverge early in the trial (−600 to −400 ms). This analysis demonstrates that none of these differences is significant, even at the level of $p < 0.2$.

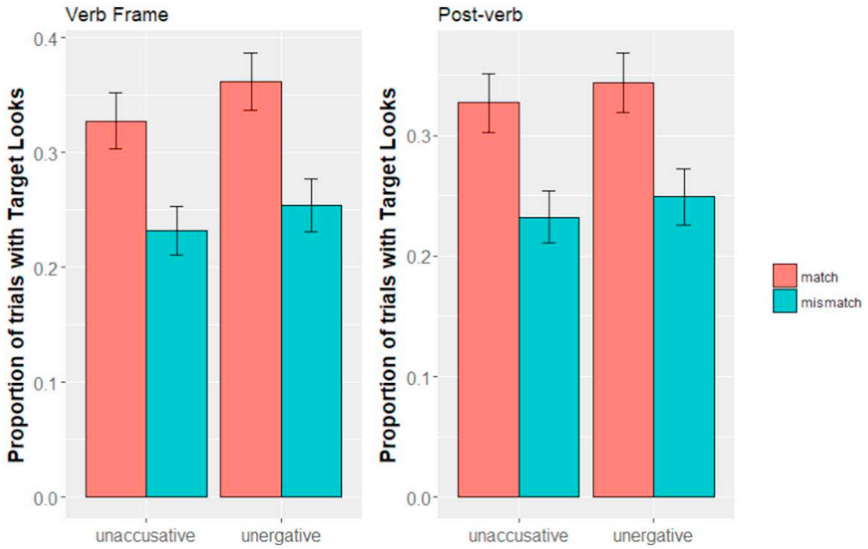

Fig. 7. Experiment 2: Proportions of trials with target looks in the verb window (left) and post-verb window (right). The error bars reflect the 95% confidence intervals.

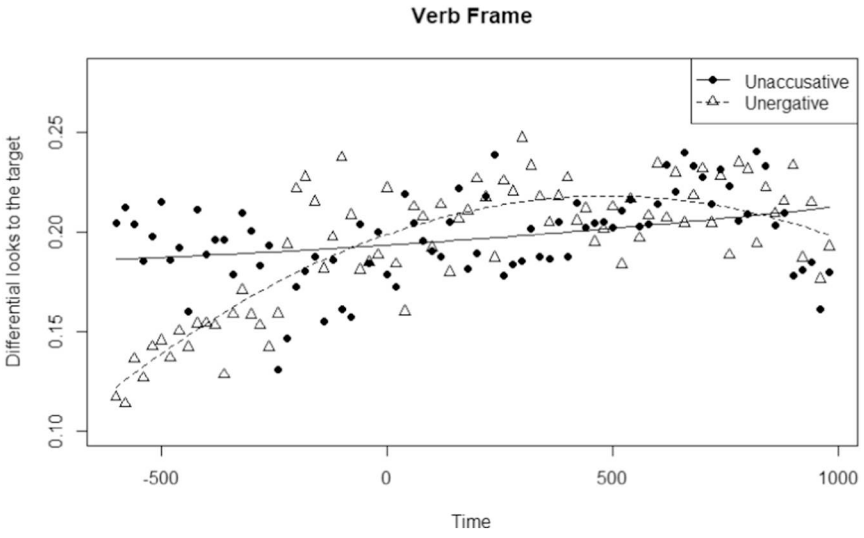

Fig. 8. Experiment 3: Fitted lines for the differential looks to the target in the two conditions. Differential looks is the difference between the proportion of looks to the target in the match condition and the proportion of looks to the target in the mismatch condition. Zero is the verb offset. The individual points correspond to the average of the dependent variable (across subjects and items) at each time point (every 20 ms).

4.4. Summary

In sum, the results of Experiment 3 closely paralleled those of the first two Experiments. First, in the growth curve analyses several of the interactions between verb condition and the time parameters had very low *p*-values, *suggesting* that the differences were highly significant. However, the pattern of these effects was different from Koring et al. (2012) and from the other studies. Since the growth curve analyses use a difference score as their dependent variable (match – mismatch) these effects are parallel to changes over time in the interaction between match and verb type in the logistic regression. We found no evidence of such an interaction in the two large time windows used in that analysis. We considered the possibility that this analysis was missing transient effects that were washed out by noise in these larger time windows. To rule this out, we ran tests on every 100 ms window in preparation for a cluster analysis, but we failed to find a single time window in which there was evidence for differential activation of the subject in the two types of verbs. Taken together these results, and those of the first two experiments, strongly suggest: 1) that there are no detectable differences in subject reactivation between the two verb types but 2) that growth curve analyses, as implemented in this paper and in Koring et al. (2012) produce false positives. In Section 5, we directly test this second claim.

5. Resampling analysis

Across our three experiments, the logistic mixed effect models and the cluster analyses did not find a difference between the unaccusative and unergative verbs, while the Growth Curve models found effects on several different terms. One might argue that the Growth Curve models can detect differences that these other models cannot, because the Growth Curve analyses model changes in fixation proportions over time, while the other models do not. Even if the match effect does not differ between the unergative and unaccusative verbs at any single moment during the trial (as the cluster analyses show), the changes in those curves might be different over time.

If this were the case, and the shape and the direction of the curves reflected specific properties of language processing (e.g. when the antecedent is reactivated, how long the reactivation/integration lasts, etc.), then we would expect these curves to have the same shape across our three experiments, and across research groups. This is not what we found. The pattern of significant effects in our growth curve analysis was different in each study and in each case different than in the Koring et al. (2012) study. This is surprising given that our three studies have exactly the same stimuli and very similar procedures. Therefore, we are left with two possibilities. The first possibility is that the growth curve models are anti-conservative and produce significant effects when there are none, perhaps due to the high correlations in fixation proportion to

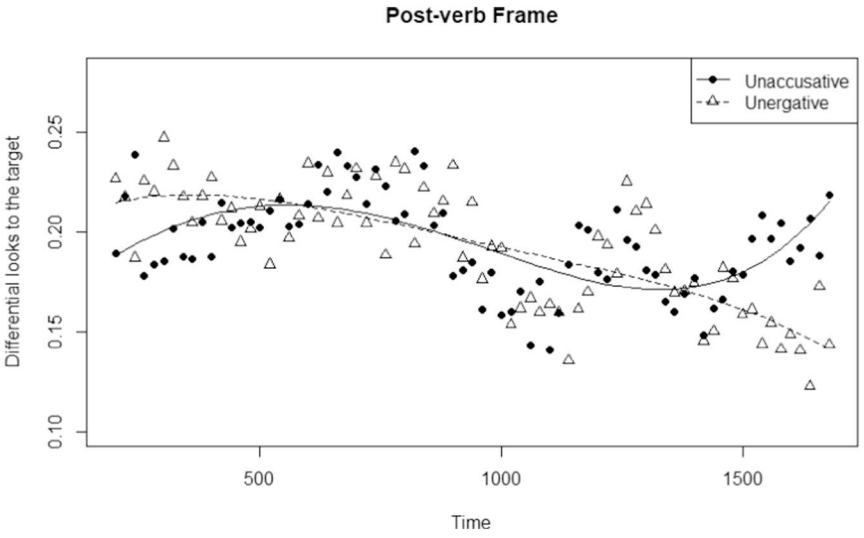

Fig. 9. Experiment 3: Fitted lines for the differential looks to the target in the two conditions. Differential looks is the difference between the proportion of looks to the target in the match condition and the proportion of looks to the target in the mismatch condition. Zero is the verb offset. The individual points correspond to the average of the dependent variable (across subjects and items) at each time point (every 20 ms).

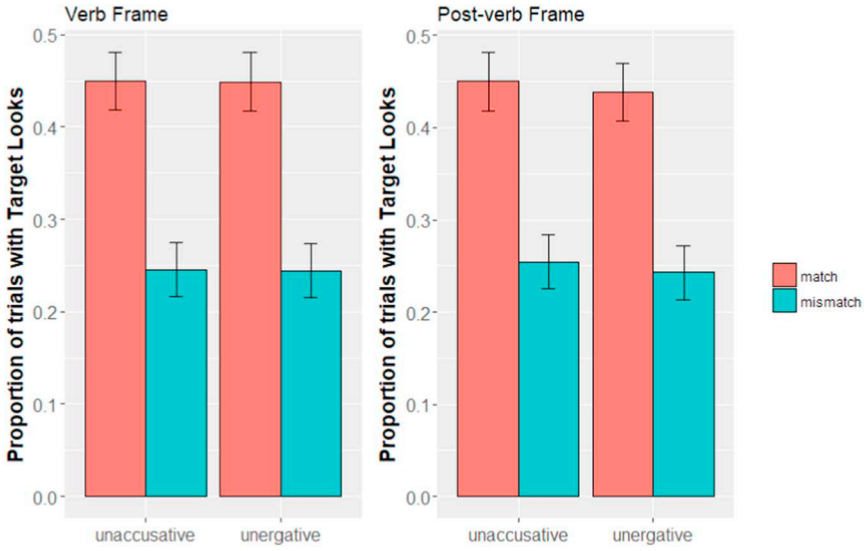

Fig. 10. Experiment 3: Proportions of trials with target looks in the verb window (left) and post-verb window (right). The error bars reflect the 95% confidence intervals.

Table 8
Results of the resampling analysis for all experiments. False positives and the p-values for alpha = 0.05 were calculated from a sample of 1000 reshufflings. Asterisks mark cases where the false positive rate is significantly greater than expected by a Fisher's exact test.

Growth curve analysis							
	Experiment 1			Experiment 2		Experiment 3	
	Term	False positives	p-Value for true alpha level of 0.05	False positives	p-Value for true alpha level of 0.05	False positives	p-Value for true alpha level of 0.05
Verb	Linear∗Condition	69%*	< 0.0001	65%*	< 0.0001	64%*	< 0.0001
	Quadratic∗Condition	50%*	< 0.0001	57%*	< 0.0001	59%*	< 0.0001
Post-verb	Linear∗Condition	64%*	< 0.0001	68%*	< 0.0001	67%*	< 0.0001
	Quadratic∗Condition	52%*	< 0.0001	53%*	< 0.0001	59%*	< 0.0001
	Cubic∗Condition	40%*	< 0.0001	44%*	< 0.0001	38%*	< 0.0001
	Quartic∗Cond	18%*	< 0.0009	33%*	< 0.0001	38%*	< 0.0001
Logistic regression							
Verb	Verb	3%	0.06	3%	0.07	3%	0.07
	Match∗verb type	8%*	0.02	7%	0.03	12%*	0.01
Post-verb	Verb	4%	0.07	5%	0.05	3%	0.07
	Match∗verb type	14%*	0.01	6%	0.04	16%*	0.008

the target between adjacent time points. The second possibility is that, the growth curve analyses are not anti-conservative, but the underlying process of reactivation does not cleanly map onto specific polynomials in these time windows and thus the verb difference, while real, shows up on different coefficients across different experiments, making it difficult to replicate the same pattern of effects twice.

To test the first possibility, we conducted a resampling analysis to empirically determine the distribution of effects for the null hypothesis given our data (similar approaches have been employed in ERP research see Piai, Dahlslätt, & Maris, 2015). The resampling procedure followed the same logic as the permutation test described in the cluster analysis section. Again, in each iteration, we randomized the condition labels (unaccusative vs. unergative) given to each item by switching half of them. The resulting data set shared two critical features of our original dataset. First, because the same label was assigned to an entire trial (rather than to individual time points), this resampling method retains the temporal dependencies within each trial which are important for time series analyses. Second, since labels were re-assigned by item (rather than being randomized by individual trial), any structure that was attributable to an individual sentence, word or picture (rather than a class of stimuli) was preserved. Thus, our resampled data retained the problematic correlational features of visual world paradigm data but instantiated the null hypothesis with respect to the two verb classes. After each reshuffling, we performed the growth curve analyses described in Section 2.3.3 on the resampled data, and noted the effects that were observed. Then we resampled again. By doing this 1000 times, for each analysis, we obtained an estimate of the expected effects of verb type under the null hypothesis. This procedure is summarized as follows:

1. Loop through all the 40 items. For each item i (a set of sentences as in Table 1), randomly decide whether to switch the condition label (unaccusative and unergative) or not by 50% chance.
2. Conduct the growth curve analyses described in Section 2.3.2, using the new randomly assigned condition labels in place of the actual condition.
3. Record the p-value for all effects of condition (the main effect and the interaction between condition and each of the time parameters).
4. Repeat Steps 1–3 1000 times.
5. Compare the distribution of p-values obtained via these simulations to the predicted distribution under the null hypothesis to determine whether the models are anticonservative.

We conducted this analysis for both time windows in all three experiments. The results are summarized in Table 8. When condition labels are randomly assigned, we should expect to find effects that are significant at the $p < 0.05$ level approximately 5% of the time, but for each individual parameter we found such effects in 18% to 69% of the reassignments. In fact, for each parameter, in each data set, the number of false positives was greater than would be expected by chance (all p's < 0.001 by a Fisher's exact test). Thus, the growth curve analyses are highly anti-conservative.

To compare this result with the performance of logistic regression models, we used a similar resampling procedure in which we shuffled the condition labels for both match and verb type and conducted the analyses described in Section 2.3.3. The results appear in Table 8. In these samples, we found p-values of < 0.05 just 3–5% of the time for the main effect of verb type and 6–16% of the time for the interaction of verb type and match. Taken at face value, these results suggest that our logistic regressions were modestly anti-conservative for the interactions but appropriately conservative for the main effect.

To get a sense of how anti-conservative these analyses are, we found, for each parameter, the 50th smallest p-value in our sample of 1000 reshufflings. These critical p-values appear in Table 8. If our models correctly captured the probability of false positive under the null hypothesis, then the p-values for these effects would be clustered near 0.05. The actual values give us a rough and ready measure of just how small a p-value would need to be in these models before we could reject the null hypothesis with an alpha of 0.05. For the growth curve analysis, these critical p-values were typically < 0.0001. In other words, these effects were > 5000 times more likely under the null hypothesis than the p-value indicates. In contrast, for the logistic regression analysis, the fiftieth p-value in the resampling set ranged from 0.07 to 0.05 and was roughly centered on 0.05 for the main effect and 0.008 or 0.02 for the interaction, suggesting that our assumptions about the null distribution in that analysis were more or less accurate.

6. General discussion

In three experiments we tested the claim that there is a fundamental difference in how unaccusative and unergative verbs are processed due to the greater syntactic complexity of unaccusatives. We found no evidence to support this claim. When we conducted simple analyses over large time windows, using logistic models, we found no differences between the verb classes. When we conducted a more fine-grained cluster analyses, we again found no differences. Curiously, however, when we conducted growth curve analyses, closely modeled on Koring et al. (2012), we found effects of verb class on the temporal parameters, in both of our time windows and in all of our studies, which *appeared* to be highly significant. But critically, these apparent effects were not consistent with those that Koring et al. (2012) reported, nor were they consistent across our three experiments. To check the validity of these analyses, we conducted a resampling analysis, using our experimental data, with verb condition labels randomly assigned. This analysis revealed that, for our data sets at least, the growth curve analyses produced p-values that were anti-conservative. In the remainder of this discussion, we address the following topics: (1) possible reasons for this high proportion of false positives in the growth curve models; (2) the rationale behind conducting a close but not exact replication; (3) the implications of these data for theories about the argument structure of intransitive verbs.

6.1. Growth curve analysis

Growth curve analyses are an attractive method for analyzing visual world paradigm data because they allow us to model changes in fixation proportion over time and thus take advantage of the rich temporal properties of this data. These analyses were first applied to visual world paradigm data by Mirman and colleagues (Mirman, Dixon et al., 2008) and appear to be gaining in popularity (Mirman & Magnuson, 2009; Kukona et al., 2011; Brown et al., 2011; Lee et al., 2013; Hadar et al., 2016; Pozzan et al., 2016; Cane et al., 2017; i.a.). As of May 22, 2018, among the research articles that cited (Mirman, Dixon et al., 2008)on Google Scholar, 111 used growth curve analysis on visual world paradigm data. As the models are typically implemented, growth curve analyses of gaze data have three limitations that may contribute to the instability that we observed. First, visual world paradigm data are highly autocorrelated over time. If a participant is looking at the target in one time sample they are highly likely to be looking at it in the next one. Koring et al. (2012), following Mirman, Magnuson et al. (2008), did not aggregate data into longer time bins or take additional modeling steps to address the problem of autocorrelation. Because we sought to replicate Koring et al.'s findings, our analyses share this limitation. Our data, like all visual world paradigm data, show high levels of autocorrelation. For example, in the verb time window of Experiment 1, we found significant evidence for autocorreltation ($p < 0.001$) by the Durbin-Watson test.

One proposed solution for the autocorrelation problem is to aggregate neighboring time points into larger time bins until the adjacent bins are no longer closely correlated (Barr, 2008; Mirman, 2014). A number of researchers have pursued this strategy, grouping their data into bins of between 50 and 200 ms before conducting growth curve analyses (Barthel, Sauppe, Levinson, & Meyer, 2016; Cane et al., 2017; Pozzan et al., 2016). How big a bin is necessary to remove autocorrelation is an empirical question which is likely to vary across studies depending on the size of the fixation region, the complexity of the display and the experimental design. We estimated the autocorrelation function in our data (ACF by (Hyndman & Khandakar, 2008)) and discovered that autocorrelation was significant up to lag of 1260 ms. When we then aggregated our data into larger bins we found highly significant autocorrelation for bins of 50, 100 and 200 ms. If this pattern is typical, it would mean that we cannot use aggregation to address the problem of autocorrelation without losing the temporal information that motivates researchers to use growth curve analyses in the first place. It seems more promising to address the problem of autocorrelation by directly modeling it instead (Pinheiro & Bates, 2000a, 2000b).

The second problem with the version of growth curve analysis, as proposed by Mirman, Dixon, and Magnuson (2008) and pursued in Koring et al. (2012), is that it averages the looks, at a given time point, for all trials in a given condition (for a given subject). This variable is then treated as a ratio variable on a linear scale. Underlyingly, however, target fixation at any given moment is binary—you are either looking at it or you're not. This creates several potential issues (see Jaeger, 2008). First, linear scales systematically distort data that are aggregates of binary measures (squeezing the extremes and stretching the center). Second, averaging across a small number of trials per cell (10 in the present study) ensures that the observed data will fall into clumps, violating the assumptions of parametric statistics. Finally, it is unclear what we are capturing at a cognitive level with this notion of continuous proportions changing over time. If the actual process is one of either looking or not (or shifting vs. staying) then what mental representation does this derived variable map onto? This set of problems can be avoided by using logistic or quasi-logistic growth curve models (see Mirman, 2014), though these have not been widely adopted in psycholinguistic studies.

The third potential problem with growth curve analyses is that there is no clear basis for determining which temporal parameters and interactions should be included in the model. Should we include quadratic, cubic and quartic parameters? What cognitive processes do we think they reflect? If our critical independent variable can enter the model in multiple ways, as an interaction with each of these temporal parameters, then we increase the odds of false positives. This risk would be minimized if we had strong and well-motivated linking hypotheses about the interpretation of these different polynomials that would allow us to make predictions and pre-specify our analyses. But in our current

knowledge state there is no reason to think that a quadratic or quartic component picks out a particular kind of process. We suspect that this problem is intrinsic to the method and not merely temporary. After all, the curve that we fit depends heavily on the decisions we make about where to begin and end our analysis, in many cases (like the present one) these choices are fairly arbitrary. This curve, as we noted above, is a derived variable that summarizes a set of underlyingly binary decisions; there is no reason to believe that the mathematically simple features of this derived measure will be the cognitively relevant ones.

The final problem with most growth curve models is that the different items (words, sentences, or pictures) are averaged together to get by participant variable that enters into the analysis. Consequently, these analyses do not model variance across items and cannot support claims about generalizations across items. This is problematic for psycholinguistic studies, where we are typically building theories about classes of words and sentences, rather than theories about particular instances (Clark, 1973). Koring et al. (2012) are not making a claim about the difference between the average of their specific collection of unergative verbs and the average of their set of unaccusative verbs, they are making a deep claim about unergatives and unaccusatives as discrete classes. Entering the items separately, and modeling their variance, is a first step in testing this more interesting claim. (The next steps would be to ensure that the verbs in the stimulus set constituted a representative sample of all verbs in the relevant class and then to determine whether the patterns of effects were not just robust across items but also categorical within each class.) In our logistic analysis, we found that the model with an item random effect is significantly better compared to a model without it ($p < 0.001$), demonstrating that the item variance matters.

It is unclear what role each of these factors played in the high false positive rate that we observed in our growth curve analyses, nor is it clear how far this problem generalizes. Until these issues are resolved, we recommend that researchers who are considering using a growth curve analysis conduct a resampling analysis (as described in Section 5) to test the empirical false positive rate for their specific model. If the model has an inflated false positive rate, researchers may want to consider conducting a cluster analysis with a permutation test instead (as described in Section 2.3.4). Cluster analysis uses small time bins and therefore retains the temporal resolution of the data, though unlike the growth curve analysis, it provides no information about the shape of the curve. Because the *p*-value for a cluster analysis is empirically determined by permutation statistics, there is no need to worry about an inflated rate of Type I errors.

There are other approaches to modeling visual world paradigm data that are emerging and seek to address the issues above. For example, Cho et al. (2018) used an autoregressive generalized linear mixed effect model to analyze visual world paradigm data. This model explicitly models the autocorrelation in the data and does not require the aggregation across subjects or items. Oleson, Cavanaugh, McMurray, and Brown (2017) analyzed visual world paradigm data by fitting individual curves and comparing groups of curves. Comparisons were done by bootstrapping and the autocorrelation problem was addressed by using a family-wise error correction.

6.2. On the diverse forms and goals of replication

As we noted in the Introduction, these experiments are close, but not exact, replications of the Koring et al. (2012) experiment. We tested the same hypothesis (that there is delayed reactivation of the subject for unergative verbs). We used the same basic paradigm (passive listening in a visual world) and the same mapping hypothesis (subject reactivation leads to eye-movements to semantic associate of the subject shortly after the verb is identified). We conceptualized and implemented our independent variables (match and verb class) in the same way. In fact, many of our verbs were direct translation equivalents. We modeled our sentences, displays and growth curve analysis on this paper.

There were three differences that make our first experiment a close replication rather than an exact replication (see Appendix 6 and LeBel et al., 2017). 1) We tested English speakers; 2) We created new stimuli (in English) that controlled for several factors that were not controlled in the original experiment; 3) We encouraged participants to attend by telling them that we would ask them questions after the study was over.

LeBel et al. (2017) note that replications exist on a continuum from exact replications, in which everything that can be controlled is identical from the stimuli to the population, to very distant conceptual replications that are linked to the original study only by a hypothesis. The goals of replication change across this continuum. Exact replications determine whether we can reproduce the same findings twice, under precisely the same conditions, but they do little to validate the hypothesis behind the study or expand the scope of generalization. Conceptual replications can provide the greatest additional support for the theory if they succeed, but, if they fail, they tell us nothing about the reproducibility of the original phenomena and little about its scope.

Le Bel and colleagues argue that close replications (like ours) are one form of direct replication (see also Schmidt, 2009). The changes that are made in conducting a close replication are typically ones that the existing theories predict should not matter if the phenomenon is stable under the description provided. For example, syntactic theories posit that the unergative and unaccusative verbs are represented in the same way across languages, and prior psycholinguistic studies argue for the presence of these processing patterns in English (Burkhardt et al., 2003; Friedmann et al., 2008). Thus, if the Koring et al. (2012) findings are attributable to the unergative and unaccusative distinction, then they should be present in English and should not disappear when the stimuli are more tightly controlled. A close replication tests whether a phenomenon is reproduceable across the range of contexts where we would expect it (across sentences, across pictures, across languages and people). Conducting a close replication was most consistent with our research goals of determining whether this pattern was stable enough to build upon. Was it consistent enough in adults to allow us to interpret data from populations, like children, that might be expected to vary in their processing (see Koring, Mak, Muders & Reuland, 2018)?

Critically, our close replication did not confirm the original hypothesis. We did not find the same data pattern as Koring et al. (2012) in any of our growth curve models. When a close replication produced null results, there are three potential explanations: 1) The replication is a false negative and the effect is real (and of the scope predicted). 2) The replication is a true negative, the original finding is a true positive, and the difference in outcomes is due to systematic differences between the studies. 3) The original finding is a false positive. We explore these hypotheses in turn and conclude that the third option best explains the set of findings to date.

When we fail to replicate a finding, it is always possible that the effect is real but the second experiment failed to detect it due to chance alone. In the present case, however, that explanation seems unlikely. First, there are three experiments, not one, all of which fail to replicate the original data pattern. Critically, across these experiments, the GCA's did not produce the pattern of effects that we might expect if the issue was inadequate power (effects that fail to reach statistical significance or just barely meet that threshold). Instead the GCA's produced effects with large t-statistics and small *p*-values that sometimes flipped in their direction across experiments. For example, in Koring et al. (2012) the critical interaction between condition and the quadratic parameter in the verb window was positive ($t = 7.89$, $p < 0.001$). If we took this at face value, we would conclude that we can confidently reject the hypothesis that unergative verbs have the same or less curvature than unaccusatives. However, in Experiment 2 we got a negative value for the same parameter ($t = -4.39$, $p < 0.001$), which taken at face value allows us to confidently reject the hypothesis that unergative verbs have the same or more curvature than unaccusatives. Clearly, both hypotheses cannot be false.

The second possibility is that the changes we made in the study

altered the pattern of effects. We made three primary changes in designing our close replication: a change in procedure (subjects were told that they would get comprehension questions), a change in the participants and their language (English vs. Dutch), and a change in the stimuli (more factors were controlled). Experiment 3 allows us to rule out the possibility that the procedural change was responsible for the difference in findings. This experiment was closely modeled on Koring et al. (2012): no comprehension questions were used or mentioned and the instructions were translated from their original experiment. In this experiment, as in the others, we found a pattern of effects in the growth curve analyses that did not match the one in Koring et al. (2012) papers. For example, there was an interaction between condition and the linear term in the verb window stemming from a very early period of activation for unaccusative verbs (counter to our hypothesis). As in the previous experiments, we did not find any interaction of verb class and match in the logistic mixed models or permutation cluster analyses.

The second major difference is that all three of our studies were conducted in English, while the Koring et al. (2012) study was conducted in Dutch. Unlike English, Dutch is a language that marks the distinction between unergative and unaccusative verbs in the auxiliary system. Auxiliaries were not used in the Koring et al. (2012) sentences, but one might wonder if speaking a language which marks this distinction could change the way in which one processes unaccusative verbs. If true, this explanation would radically change how we think about unaccusativity. Generative theories propose that the syntactic distinction between unergatives and unaccusatives reflects deep properties of the syntax-semantics interface and thus is present in all languages, in the same form, regardless of morpho-syntactic marking.

The possibility of a cross-linguistic difference, however, has been explored and rejected, both by our group (Huang, van Hemert, van Hout, & Snedeker, in prep) and by Koring and van de Koot (2018). Koring and van de Koot (2018) conducted a study in English that was closely parallel to Koring et al. (2012). They analyzed their results using GCA's and found two effects in English that they argued were functionally parallel to those in Dutch: a positive quadratic component in a very early window and a positive quartic component in the post-verb window.[4] They concluded that in both languages unergative subjects are reactivated early and unaccusative subjects are reactivated later. Our Dutch study (Huang et al., in prep) closely paralleled Experiment 2 in the present paper: the stimuli were controlled for frequency of the verbs, the animacy of the subjects, the plausibility of the verb given the subject and the naturalness of the sentences. Participants were told that they would hear some sentences and look at some pictures. During the study, they were asked occasional questions about what the stimuli. We found a pattern of results that mirrored the studies reported here: a) many significant effects of verb type in the growth curve analysis; b) the absence of any significant interaction between match and verb type in the logistic regressions (in any time window); and c) a high rate of false positives for the GCA's in the resampling analyses. Thus, there is no reason to think that the divergence between our conclusions and those of Koring et al. (2012) reflect a difference between English and Dutch.

The final change that we made in our replications was that we control for a greater number of potential confounds, including imageability of the verbs, the codability of the target image, the complexity of the critical sentence after the verb, the plausibility of the sentence, and how plausible the verb is given the subject noun. It is possible that these changes had an effect on the outcome. We chose not to explore this via an additional experiment. If we built these confounds into our stimuli and we replicated the data pattern, the new finding would not be a useful basis for further science, though it might assuage our curiosity. Fortunately for us, there are two strong reasons to think that the confounds alone were not attributable for the differences between our study and theirs. First, there is a more plausible explanation that is strongly supported by our simulations, as we will see below. Second, these confounds were also presumably present in Koring et al., (2018), a developmental study in Dutch which used very similar stimuli. But the data from the adults in that study patterns quite differently. In the 2018 study, there is a larger match effect for unaccusative verbs than for unergatives in verb window resulting in a difference in the intercepts that is in the opposite direction of the difference found in the 2012 paper (and in the opposite direction of what we would predict given the greater imageability of unergative verbs).

The final, and the simplest, explanation is that the results of Koring et al. (2012) are a false positive attributable to the statistical technique they employed. In our resampling analyses we learned that when we randomly shuffled the verb condition labels (arbitrarily labeling verbs as unaccusative or unergative) we would get highly significant results on each of the relevant parameters roughly half the time. Since the analyses each had two to four parameters capturing differences in the verbs, this meant that most of the random models produced false positives. In other words, if we assume that the Koring et al. (2012) data is similar, then the findings presented in that paper are pretty much what we would expect if there were no differences between the two verb classes. All research runs the risk of producing inadvertent false positives. Typically, our *p*-values provide some guidance about the extent of that risk. In the case of growth curve models applied to eye-gaze data, these p-values are inaccurate, suggesting a much higher degree of certainty than is warranted. Critically, however, no one using growth curve models (prior to our study) had any reason to question these statistics. Thus, it is possible, even likely, that there are other false positives in the psycholinguistic literature.

6.3. On the dichotomy between unaccusative and unergative verbs

Our results also bear on the question of whether there are processing differences between unergative and unaccusative verbs. We can think about the question in a narrow way: What should we conclude about the findings in Koring et al. (2012) given this data? But we can also think about it in a broader way: What should we conclude about the processing of unaccusatives given the psycholinguistic studies to date? As we noted above, the most parsimonious explanation for the discrepancy between our conclusions and those of Koring et al. (2012) is that the original finding was a false positive due to the use of an analysis method that is highly anti-conservative. This hypothesis would be consistent with: a) the instability in the pattern of effects in the growth curve models across the studies done by Koring and her colleagues; b) the instability in the pattern of effects in the growth curve models across the studies conducted by our group; c) the absence of effects in the logistic mixed models and resampling analyses in the present paper; d) the high rates of false positives that we observed for the growth curve models in the resampling analysis (Section 5).

If this third explanation is correct, then there is no evidence from the visual world paradigm for processing differences between unaccusative and unergative verbs. There are, however, a handful of experiments using other methods that *seem* to show such differences. Critically, as we noted in the introduction, two studies using the cross-model lexical decision priming paradigm (Burkhardt et al., 2003; Friedmann et al., 2008) have found differences in the time course of argument reactivation. To recap, both studies found reactivation of unaccusative subjects about 700 ms after the verb, with no reactivation for unergatives in this time window. These measures rely on semantic priming mediated by the subject noun. The present study relies on the same cognitive mechanism (measured in a different way). If these prior studies provide a stable insight into processing, we should have seen a

[4] We based our study and analyses on the Koring et al. (2012) study because the Koring and van de Koot (2018) paper was not available at the time we collected this data. The Koring and van de Koot study cannot be taken as independent evidence for the findings of Koring et al. (2012). The early unergative reactivation effect in is found in a different time window which was selected post hoc based on visual inspection of the data.

divergence in the two verb classes around 700 ms after the verb offset. But we did not. We can rule out the possibility that effects in the visual world paradigm are simply earlier or later: a fine-grained analysis over a large time frame (i.e. our cluster analysis in Sections 2.3.4, 3.3.3, and 4.3.3) did not show any differences between these two types of verbs from 600 ms before to 2000 ms after the verb offset. Therefore, it is unlikely that we missed the the critical moment for finding this effect. Another possibility is that cross-modal lexical priming is simply more sensitive than the visual world paradigm. We think this is unlikely; like the cross modal studies, we found large effects of semantic relatedness at the subject position, demonstrating that the visual world paradigm is also quite sensitive to the semantic priming.

We believe that the differences between these studies are best explained by looking at the potential confounds in the Friedmann et al. (2008) and Burkhardt et al. (2003) stimuli. The cross-modal priming studies did not equate their unaccusative and unergative stimuli for the full set of factors that we controlled for. For example, neither of the two studies reports controlling for the imageability of the verbs and neither study holds constant the sentence frames across the two verb classes (see Section 1 for more details). If less imageable verbs lead to slower argument reactivation (see e.g., Paivio, 1991) this could account for the differences observed in these studies.

The other psycholinguistic studies of unergative and unaccusative verbs use fMRI paradigms (Agnew et al., 2014; Shetreet, Friedmann, & Hadar, 2010). Both studies reported differences between sentences with these two types of verbs, though it is unclear how these patterns relate to subject reactivation. Critically, the observed patterns of activation were different in the two studies. Shetreet et al. (2010) found an activation that was specific to unaccusative sentences compared to unergative and transitive sentences. Agnew et al. (2014) "failed to find any activation that is specific to the neural processing of sentences with unaccusative verbs" instead unaccusative verbs patterned with transitives. Neither paper mentions controlling for the imageability of the verb. Both studies used different carrier sentences for the unergative and unaccusative verbs introducing potential confounds. Thus the observed effects could reflect uncontrolled properties of the items that were used rather than true (but variable effects) of verb type. In sum, the psycholinguistic studies do not provide clear evidence that unaccusatives are processed differently than unergatives. While future studies, using tighter controls and more robust statistics, may produce such evidence, for now we must remain agnostic.

What does this mean for the unaccusative hypothesis more broadly? The failure to find processing effects in no way excludes the possibility that there is an underlying syntactic difference between unaccusative and unergative verbs. If processing time largely reflects predictability and frequency rather than structural complexity (Levy, 2008), then online measures would be the wrong place to look for answers to these syntactic questions. For example, perhaps the syntax of the unaccusative verb is a tad more complex, but this complexity is built into the lexical entry and adds no measurable cost to retrieval time.[5]

In the absence of any strong processing data, support for the Unaccusative Hypothesis must come from its ability to provide unique predictions about the syntactic distribution of verbs across a range of languages. Linguists have long argued that there is a cluster of syntactic phenomena that distinguish unaccusatives from unergatives and suggest that unaccusative subjects are close kin to transitive objects (Perlmutter, 1978; Burzio, 1981, 1986; Rosen, 1984; Levin & Rappaport, 1986; i.a.). Critically, the Unaccusative Hypothesis predicts that these diagnostics should divide verbs cleanly into the same two categories across tests and across languages. However, many linguists have noted that each of the diagnostics seems to pick out a somewhat different class of verbs (Zaenen, 1993; Levin & Rappaport, 1995; Sorace, 2000; Alexiadou, Anagnostopoulou, & Everaert, 2004; Deal, 2009; i.a.). These subclasses are usually described in terms of their meanings (e.g. change of state, change of location, etc.). One plausible explanation for these patterns is that the natural taxonomy of the intransitive verbs is not based on a two-way syntactic distinction between unergatives and unaccusatives. Instead there could be many distinct syntactic subclasses. This is the pattern that we would expect if each syntactic test taps into a different aspect of the verb's semantic structure, or underlying meaning (Dowty, 1991; Van Valin, 1990; Zaenen, 1993). On such an account, we would have no reason to expect systematic differences in the processing of so-called unaccusative and unergative verbs, since each of these categories would contain a mix of verbs with different semantic structures.

[5] We thank an anonymous reviewer for point out this possibility.

CRediT authorship contribution statement

Yujing Huang: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Jesse Snedeker:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

Acknowledgements

This work has been funded by Harvard Mind Brain Behavior Interfaculty Initiative and the Institute for Quantitative Social Science at Harvard. We thank Liz Chalmers, Katherine Farnsworth, Alissa Horsung for assisting the data collection. We thank our six anonymous reviewers, the associate editor, Hugh Rabagliati and audience at AMLaP 2016 and Cogsci 2016 for their comments and suggestions. We also thank members of Snedeker lab for the help and discussion.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104251.

References

Abraham, W. (1986). Unaccusatives in German. *Groninger Arbeiten zur Germanistischen Linguistik, 28*(1), 72.

Agnew, Z. K., van de Koot, H., McGettigan, C., & Scott, S. K. (2014). Do sentences with unaccusative verbs involve syntactic movement? Evidence from neuroimaging. *Language, Cognition and Neuroscience,* 1035–1045.

Alexiadou, A., Anagnostopoulou, E., & Everaert, M. (2004). *The unaccusativity puzzle: Explorations of the syntax-lexicon interface. Vol. 5*. Oxford University Press on Demand.

Altmann, G. T., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. *The interface of language, vision, and action: Eye movements and the visual world* (pp. 347–386). .

Baker, M. C. (1988). *Incorporation: A theory of grammatical function changing.* Chicago, Illinois: The University of Chicago Press.

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474.

Barthel, M., Sauppe, S., Levinson, S. C., & Meyer, A. S. (2016). The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm. *Frontiers in psychology, 7*.

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

Bever, T. G., & Sanz, M. (1997). Empty categories access their antecedents during comprehension: Unaccusatives in Spanish. *Linguistic Inquiry,* 69–91.

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers, 35*(1), 158–167.

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (BOSS)

phase II: 930 new normative photos. *PLoS One, 9*(9), e106953.
Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin & Review, 18*(6), 1189–1196.
Burkhardt, P., Piñango, M. M., & Wong, K. (2003). The role of the anterior left hemisphere in real-time sentence comprehension: Evidence from split intransitivity. *Brain and Language, 86*(1), 9–22.
Burzio, L. (1981). *Intransitive verbs and Italian auxiliaries.* Doctoral dissertationMassachusetts Institute of Technology.
Burzio, L. (1986). *Italian syntax: A government-binding approach. Vol. 1*. Springer Science & Business Media.
Cane, J., Ferguson, H., & Apperly, I. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 591–610.
Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of experimental psychology: Learning, memory, and cognition, 30*(3), 687.
Chierchia, G. (2004). A semantics for unaccusatives and its syntactic consequences. *The unaccusativity puzzle,* 22–59.
Cho, S. J., Brown-Schmidt, S., & Lee, W. Y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: An application to intensive binary time series eye-tracking data. *Psychometrika,* 1–21.
Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335–359.
Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990–present. Available online at http://corpus.byu.edu/coca/.
Deal, A. R. (2009). The origin and content of expletives: Evidence from "selection". *Syntax, 12*(4), 285–323.
Dowty, D. (1991). Thematic proto-roles and argument selection. *Language,* 547–619.
Friedmann, N., Taranto, G., Shapiro, L. P., & Swinney, D. (2008). The leaf fell (the leaf): The online processing of unaccusatives. *Linguistic inquiry, 39*(3), 355–377.
Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17*(8), 684–691.
Hadar, B., Skrzypek, J. E., Wingfield, A., & Ben-David, B. M. (2016). Working memory load affects processing time in spoken word recognition: Evidence from eye-movements. *Frontiers in Neuroscience, 10*.
Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid linguistic ambiguity resolution in young children with autism spectrum disorder: Eye tracking evidence for the limits of weak central coherence. *Autism Research, 8*(6), 717–726.
Haider, H., & Rindler-Schjerve, R. (1987). The parameter of auxiliary selection: Italian-German contrasts. *Linguistics, 25*(6), 1029–1056.
Hartshorne, J. K., Nappa, R., & Snedeker, J. (2015). Development of the first-mention bias. *Journal of Child Language, 42*(2), 423–446.
Hoekstra, T., & Mulder, R. (1990). Unergatives as copular verbs; locational and existential predication. *The linguistic review, 7*(1), 1–80.
Huang, Y., van Hemert, A., van Hout, A., & Snedeker, J. (2020). Are unaccusatives and unergatives processed differently in Dutch? – A Visual World Paradigm study. *Unpublished manuscript.* (In prep).
Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 26*(3), 1–22.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.
Kipper, L., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation Journal, 42*(1), 21–40.
Koring, L., Mak, P., & Reuland, E. (2012). The time course of argument reactivation revealed: Using the visual world paradigm. *Cognition, 123*(3), 361–379.
Koring, L., & van de Koot, H. (2018). Processing questions. Paper presented at De Grote Taaldag, Utrecht, The Netherlands. Retrieved from https://www.loeskoring.net/m/publications/GTDWebversion.pdf.
Koring, L., Mak, P., Mulders, I., & Reuland, E. (2018). Processing intransitive verbs: how do children differ from adults? *Language Learning and Development, 14*(1), 72–94.
Kukona, A., Fang, S.-Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition, 119*(1), 23–42.
LeBel, E. P., Berker, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology, 113*, 254–261.
Lee, C.-l., Middleton, E., Mirman, D., Kalénine, S., & Buxbaum, L. J. (2013). Incidental and context-responsive activation of structure-and function-based action features during object identification. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 257.
Legendre, G. (1989). Unaccusativity in French. *Lingua, 79*(2), 95–164.
Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* University of Chicago press.
Levin, B., & Rappaport, M. (1986). The formation of adjectival passives. *Linguistic inquiry,* 623–661.
Levin, B., & Rappaport, M. (1995). *Unaccusativity: At the syntax-lexical semantics interface.* Cambridge, Massachusetts: MIT Press.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190.
Mirman, D. (2014). *Growth curve analysis and visualization using R (Chapman & Hall/CRC the R series)*. Boca Raton: CRC Press.
Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*(4), 475–494.
Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition, 37*(7), 1026–1039.
Mirman, D., Magnuson, J. S., Estes, K. G., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition, 108*(1), 271–280.
Momma, S., Slevc, L. R., & Phillips, C. (2018). Unaccusativity in sentence production. *Linguistic Inquiry, 49*(1), 181–194.
Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (2017). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research, 26*(6), 2708–2725.
Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology, 76*(1p2), 1.
Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie, 45*(3), 255.
Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition, 108*(1), 155–184.
Perlmutter, D. M. (1978). Impersonal passives and the unaccusative hypothesis. *Proceedings of the annual meeting of the Berkeley Linguistics Society. Vol. 4*.
Piai, V., Dahlslätt, K., & Maris, E. (2015). Statistically comparing EEG/MEG waveforms through successive significant univariate tests: How bad can it be? *Psychophysiology, 52*(3), 440–443.
Pinheiro, J., & Bates, D. M. (2000a). *Mixed-effects models in S and S-PLUS (Statistics and computing)*. New York: Springer.
Pinheiro, J. C., & Bates, D. M. (2000b). *Linear mixed-effects models: Basic concepts and examples. Mixed-effects models in S and S-Plus.* 3–56.
Pozzan, L., Gleitman, L. R., & Trueswell, J. C. (2016). Semantic ambiguity and syntactic bootstrapping: The case of conjoined-subject intransitive sentences. *Language Learning and Development, 12*(1), 14–41.
R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for statistical computing, Vienna, Austria*. URL https://www.R-project.org/.
Reuter, T., Feiman, R., & Snedeker, J. (2018). Getting to no: Pragmatic and semantic factors in two-and three-year-olds' understanding of negation. *Child Development, 89*(4), e364–e381.
Rosen, C. (1984). The interface between semantic roles and initial grammatical relations. *Studies in relational grammar, 2*(38–77).
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100.
Shetreet, E., Friedmann, N., & Hadar, U. (2010). The neural correlates of linguistic distinctions: Unaccusative and unergative verbs. *Journal of Cognitive Neuroscience, 22*(10), 2306–2315.
Simpson, J. (1983). Resultatives. In L. Levin, M. Rappaport, & A. Zaenen (Eds.). *Papers in lexical-functional grammar* (pp. 143–157). .
Sorace, A. (2000). Gradients in auxiliary selection with intransitive verbs. *Language,* 859–890.
Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Bates, E., et al. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language, 51*(2), 247–250.
Van Valin, R. D., Jr. (1990). Semantic parameters of split intransitivity. *Language,* 221–260.
Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 1.
Zaenen, A. (1988). *Unaccusative verbs in Dutch and the syntax-semantics interface. Vol. 123.* CSLI/SRI International.
Zaenen, A. (1993). Unaccusativity in Dutch: Integrating syntax and lexical semantics. *Semantics and the Lexicon* (pp. 129–161). Netherlands: Springer.