Alignment between Thematic Roles and Grammatical Functions Facilitates Sentence Processing

Evidence from Experiencer Verbs

Michael Wilson
University of Delaware

Brian Dillon
University of Massachusetts Amherst

Abstract

How does grammatical markedness affect processing? Previous work has studied this extensively in the domain of experiencer verbs, by examining the question of alignment between thematic role and grammatical function hierarchies. Existing evidence is consistent with multiple accounts, including an experiencer-first preference or an experiencer-subject preference. We conducted two experiments to disentangle these effects for experiencer verbs, using self-paced reading with comprehension questions and speeded grammaticality judgments. Our results revealed a clear experiencer-subject preference, and no experiencer-first preference in the processing of English. These results are most consistent with a view where grammatical constraints cannot be reduced to more general cognitive constraints.

Keywords

sentence processing; self-paced reading; thematic relations; argument structure; lexical semantics; Bayesian data analysis

1. Introduction

To understand a sentence, comprehenders must link a verb's arguments with their thematic roles. What factors are relevant to how this linking process occurs in real-time language use? One factor appears to be how canonical the alignment between a thematic role and its position is. Non-canonical mappings—for instance, as found in passive sentences—famously pose difficulties in language comprehension (e.g., Ferreira 2003) and acquisition (e.g., Nguyen & Pearl 2018, 2021). But exactly what makes non-canonical sentences difficult to comprehend remains under debate. In this paper we contribute to this body of work by exploring the role of alignment in comprehension with two experiments examining subclasses of experiencer verbs.

Contact information: Michael Wilson, mawilson@udel.edu, 125 E Main St, Newark, DE, 19702 (corresponding author); Brian Dillon, brian@linguist.umass.edu. This research was funded by the University of Massachusetts Amherst. The data and code used to conduct the research described in this paper are available at osf.io/jq62t.

We extend our thanks to Nazbanou Nozari, two anonymous *Journal of Memory and Language* reviewers, and one non-anonymous *Journal of Memory and Language* reviewer (Shravan Vasishth) for their insightful, critical, and constructive comments. For additional valuable discussion and feedback, we thank Shota Momma, Lyn Frazier, Elsi Kaiser, Monica Do, Emma Nguyen, Seth Cable, and Rong Yin, as well as the University of Massachusetts Amherst Psycholinguistics and Syntax Reading Groups, the members of the University of Massachusetts Amherst Cognitive Science of Language Lab, and attendees of the 2020 AMLaP conference. We also thank Barbora Hlachova for assistance in running participants for the suspended eye-tracking study presented in Appendix E. We thank Bob Frank for providing access to the computational resources without which we could not have run our statistical analyses. Any errors that remain are our own.

Experiencer verbs are of crucial importance in the study of thematic role linking processes. This is because two subclasses of experiencer verbs, subject experiencer verbs and object experiencer verbs, invoke precisely the same (or at least very similar) sets of thematic roles: experiencer and theme. However, these verbs link these thematic roles to grammatical functions in precisely the opposite way, as shown in (1).

(1) a. $[\text{John}]_{\text{experiencer}}$ fears $[\text{ghosts}]_{\text{theme}}$. (Subject experiencer)

b. $[\text{Ghosts}]_{\text{theme}}$ frighten $[\text{John}]_{\text{experiencer}}$. (Object experiencer)

Subject experiencer verbs are often considered to present a more canonical mapping between thematic role and position—for reasons we detail below, we refer to this as a mapping where there is *alignment* between thematic roles and grammatical functions. In contrast, active object experiencer verbs are *misaligned*, presenting a non-canonical mapping. As we detail below, previous work has found that sentences like (1a) are easier to process, understand, and produce than sentences like (1b). We aim to disentangle two competing explanations of this fact.

The first potential explanation, which we will call **experiencer-first**, holds that aligned configurations like (1a) are easier to process because the linear order of this type of sentence corresponds to the salience or accessibility of the arguments (or perhaps, their thematic roles) (Bader et al. 2017; Gattei et al. 2022, 2015a, 2017, 2015b; Gennari & MacDonald 2008, 2009; Temme & Verhoeven 2016; Verhoeven 2014). In this view, sentences like (1a) are easier to process than (1b) because the argument bearing the more salient or accessible thematic role (experiencer is more salient than theme) occurs first. More accessible thematic roles are more easily activated, and thereby preferentially linked to earlier arguments for various reasons relating the cognitive salience of such roles (such as animacy, frequency of occurrence (e.g., Gennari & MacDonald 2009), etc.). This is similar to a well established agent-first preference (e.g., Bickel et al. 2015; Christianson & Cho 2009; Ferreira 2003; Gómez-Vidal et al. 2022; Hammerly 2020; Hammerly et al. 2022; Huber et al. 2023; Isasi-Isasmendi et al. 2023 a.m.o.), with the only difference being the specific thematic role(s) involved.

The second potential explanation, which we will call **experiencer-subject**, holds that aligned configurations like (1a) are easier to process because the argument with the higher ranked thematic role is mapped to a higher structural position than the argument with the lower ranked thematic role. Grammatical constraints favor such mappings over ones where the opposite pattern holds (as in (1b)). Notably, these constraints do not make direct reference to surface-level information like linear order, and do not necessarily justify themselves straightforwardly on the basis of more general cognitive principles (such as an activation "race" where more highly activated candidates or thematic role linkings are retrieved first; e.g., MacDonald 2013). These constraints are justified on grammar-internal grounds, as they have been invoked to explain systematic regularity in how arguments are mapped to structural positions across verbs and languages (e.g., Baker 1988; Belletti & Rizzi 1988; Perlmutter & Postal 1984).

We aim to better understand the source of the difficulty associated with misaligned configurations. To this end, we conducted two experiments examining the processing of experiencer verbs in active and passive voice. These experiments were designed so that the two competing explanations for alignment effects would make precisely the opposite predictions regarding which configurations would be found more difficult.

In English, it is impossible to determine which of these preferences may apply by examining simple monoclausal sentences, as in (1). This is because these sentences entirely confound linear

order and grammatical function: subjects go first, and objects second. Any evidence based on such sentences cannot address the question of which preference is active—or, more precisely, the relative strength of these preferences. This is because both an experiencer-first preference and an experiencer-subject preference would favor the same linking structure. If both preferences are active, they would reinforce each other, and it would be impossible to tell whether one preference contributes more to any observed effect. Assuming only one of the preferences is active, it would be impossible to tell which of the two possibilities it is.

To our knowledge, all previous studies on the linking preferences of experiencer verbs in English fail to distinguish these two possibilities. Though authors have interpreted findings of difficulty for misaligned configurations as evidence for one or the other of these effects variously (e.g., experiencer-subject: Do & Kaiser 2019, 2022; Ferreira 1994, a.o.; experiencer-first: Bader et al. 2017; Temme & Verhoeven 2016; Verhoeven 2010, 2014, 2015, a.o.), studies on English experiencer verbs have centered on the use of simple monoclausal sentences where linear precedence is confounded with grammatical function. Thus, previous results for English are largely predicted by either sort of preference, or by both preferences favoring the same linking structure. Results from studies on non-English languages, where linear order and grammatical function are easier to deconfound (e.g., German, Spanish) have found effects that are most consistent with the experiencer-first preference (Bader et al. 2017; Bornkessel et al. 2003; Gattei et al. 2022, 2015a, 2017, 2015b; Temme & Verhoeven 2016; Verhoeven 2010, 2014, 2015), though some results suggest the existence of a potentially separate experiencer-subject preference as well (e.g., Verhoeven 2014).

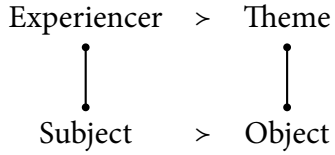
We begin with a discussion of the linking behavior of experiencer verbs from the formal syntax and semantics literature to make precise the notion of alignment relevant to the experiencer-subject preference. Then, we review evidence from the psycholinguistic literature that bears on the experiencer-first and experiencer-subject preferences, before turning to our experiments. The results of these experiments are most consistent with a strong experiencer-subject preference, and provide no support for an experiencer-first preference being active in English. We suggest that the source of alignment effects may reside in grammatical markedness constraints that constitute thematic role linking rules (e.g., Baker 1988; Perlmutter & Postal 1984).

1.1 Linking and alignment

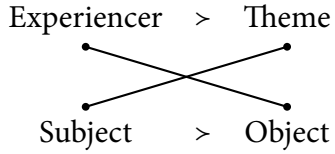
According to one line of thinking, the mapping between grammatical functions and thematic roles is governed by a set of very general regularities (e.g., Baker 1988; Perlmutter & Postal 1984). With regards to experiencer verbs in particular, Belletti & Rizzi (1988)'s influential proposal has proved a fount of inspiration for much syntactic, semantic, and psycholinguistic work. They assume that there are universal linking rules that make reference to a thematic role hierarchy. These linking rules describe how arguments bearing particular thematic roles are realized syntactically via **alignment**: roles that are higher on the thematic role hierarchy are linked to higher grammatical functions (or perhaps syntactic positions), whether these positions are defined in absolute (e.g., Baker 1988) or relative (e.g., Perlmutter & Postal 1984) terms. We can assume the subset of the thematic hierarchy relevant to experiencer verbs is *Experiencer* > *Theme* (cf. Baker 1988, 1989; Belletti & Rizzi 1988; Fillmore 1968; Givón 1984; Grimshaw 1990; Jackendoff 1972, 2002; Perlmutter & Postal 1984; Rappaport & Levin 1988; Van Valin 1990), and the relevant grammatical function hierarchy is *Subject* > *Object*.

Belletti & Rizzi (1988) proposed both subject and object experiencer verbs underlyingly display the same linking properties. However, the underlying structure associated with object experiencer verbs is not transparently mapped onto a surface string, instead undergoing a passive-like operation that reverses the order of the arguments. This leads to object experiencer verbs displaying misaligned linking behavior on the surface.

(2) a. Subject experiencer verbs (aligned):



b. Object experiencer verbs (misaligned):

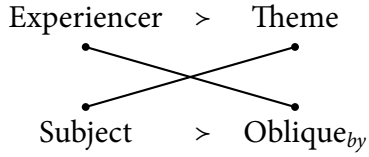


This makes it clear that in active transitive sentences, subject experiencer verbs display aligned linking behavior, and object experiencer verbs display misaligned linking behavior. In our diagrams, misalignment is visually apparent as a crossing of the lines that link each thematic role to the grammatical function that realizes it.

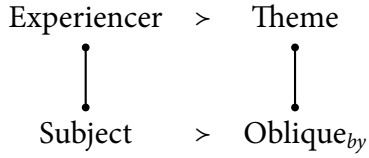
1.2 Psycholinguistic background

While Belletti & Rizzi (1988) do not discuss how alignment applies to passives, the lion's share of research on misalignment in language use has focused on the difficulties posed by the passive voice in acquisition, comprehension, and production (e.g., Balogh & Grodzinsky 1996; Cupples 2002; Do & Kaiser 2022; Ferreira 1994, 2003; Grodzinsky 1995; Grodzinsky et al. 1991; Nguyen & Pearl 2019, 2021; Piñango 2000). To bridge this gap, we can treat thematic role alignment as describing properties of not only canonical (i.e., active transitive) sentences, but also other kinds of sentences. To show how this works, consider the linking diagrams from (2), modified for passive uses of different subclasses of experiencer verbs, shown in (3).

(3) a. Subject experiencer passives (misaligned):



b. Object experiencer passives (aligned):



As (2–3) make apparent, comparisons between active and passive uses of verbs have proven quite fruitful because they display precisely the opposite linking behavior in their surface syntax.

The non-canonical linking between grammatical function and thematic role is plausibly one source of difficulty for passive constructions, helping provide insight into the nature of difficulties associated with the comprehension of passive sentences (Ferreira 2003) and the factors that ameliorate these difficulties (Slobin 1966). This suggests a broader hypothesis: that (surface) aligned configurations should be easier to produce and process than (surface) misaligned configurations (cf. Bornkessel & Schlesewsky 2006; Bornkessel-Schlesewsky & Schlesewsky 2009; Do & Kaiser 2019, 2022; Ferreira 1994; Hammerly 2020; Hammerly et al. 2022; Wagers & Pendleton 2016; Wagers et al. 2018). Indeed, much previous psycholinguistic work has defended the hypothesis that alignment facilitates production and processing by investigating its effects in domains other than experiencer verbs (e.g., Bornkessel & Schlesewsky 2006; Bornkessel-Schlesewsky & Schlesewsky 2009; Hammerly 2020; Hammerly et al. 2022; Traxler et al. 2002; Wagers & Pendleton 2016).

For instance, Bornkessel & Schlesewsky (2006) and Bornkessel-Schlesewsky & Schlesewsky (2009) proposed the extended Argument Dependency Model (eADM), which is intended to capture effects of alignment across multiple grammar-internal and general cognitive hierarchies simultaneously. With regards to thematic role assignment, they propose that comprehenders predict thematic roles for arguments greedily, and prefer to link arguments encountered earlier to higher thematic roles, most consistent with an experiencer-first preference. Upon reaching the verb, its thematic roles may override the predicted ones, which can lead to reanalysis. This reanalysis is costly particularly when it requires participants to assign a relatively lower ranked thematic role to an argument already predicted to bear a relatively higher ranked thematic role (see Bornkessel et al. 2002, 2003).

Several studies have examined effects of alignment in the processing and production of experiencer verbs in English, and found evidence consistent with this view. Cupples (2002) provides evidence from plausibility judgments that misaligned configurations led to greater processing and comprehension difficulty, though these effects were not consistent across her three experiments. However, an earlier study with a similar design failed to yield clear evidence for an effect of alignment (Carrithers 1989). But the designs of these studies leave open the possibility that the relevant notion of alignment might be related to either the experiencer-first or the experiencer-subject preference, since these studies used simple monoclausal sentences in English, where both preferences predict greater relative difficulty in the same conditions.

Gennari & MacDonald (2008) showed that object relative clauses (ORCs) with active object experiencer verbs were read more slowly than subject relative clauses (SRCs) with passive object experiencer verbs, with accuracy similarly affected. Sample items for the relevant conditions are in (4).

(4) a. The director that the movie pleased had received a prize. (active ORC)
b. The director that was pleased by the movie had received a prize. (passive SRC)

However, there is a potential structural confound in their design (i.e., comparing ORCs to SRCs). Manouilidou & de Almeida (2013) found that object experiencer verbs produced longer RTs at a pre-verbal adverb and the verb, though subject experiencer verbs produced longer RTs at the determiner following the verb. Thus, both active object experiencer verbs and active subject experiencer verbs led to greater difficulty, albeit at different points in the sentence. Other studies using a variety of measures including SPR, eye-tracking while reading, and MEG have generally found results consistent with effects of alignment (Brennan & Pylkkänen 2010; Paolazzi 2018; Paolazzi et al. 2019, 2021).

Converging evidence for effects of alignment come from studies of neurologically impaired and

non-adult populations. Several studies have shown that patients with Broca's aphasia generally display greater difficulty with misaligned configurations compared to aligned configurations with experiencer verbs (Balogh & Grodzinsky 1996; Beretta & Campbell 2001; Grodzinsky 1995; Grodzinsky et al. 1991; Piñango 2000; Thompson & Lee 2009; but see Black et al. 1991). Furthermore, studies of language acquisition have shown that children's comprehension, production, and age of acquisition of passives of experiencer verbs is consistent with alignment effects, with object experiencer passives being acquired relatively earlier than subject experiencer passives (Ambridge et al. 2021; Bidgood et al. 2020; Crain et al. 2009; Fox & Grodzinsky 1998; Gordon & Chafetz 1990; Hirsch & Wexler 2006; Liter et al. 2015; Maratsos & Abramovitch 1975; Maratsos et al. 1985; Nguyen 2015; Nguyen & Pearl 2018, 2019, 2021; O'Brien et al. 2006; Orfitelli 2012; Pearl & Sprouse 2019, 2021; Perovic et al. 2014; Pinker et al. 1987; de Villiers & de Villiers 1973; but see Messenger et al. 2012 and Hartshorne et al. 2015). Some of these studies included adult control groups that displayed effects consistent with the thematic-grammatical hypothesis in production and comprehension as well (e.g., Ambridge et al. 2021; Bidgood et al. 2020; Nguyen 2021).

Effects of alignment have also been observed in sentence production tasks. Ferreira (1994) found that speakers were relatively more likely to produce simple monoclausal sentences with aligned configurations for agent-theme, subject experiencer, and object experiencer verbs (see also Dröge et al. 2014 for similar results in Italian), though these effects were strongest when arguments differed in animacy. Altmann & Kemper (2006) (see also Altmann et al. 2004) found similar results. Do & Kaiser (2019, 2022) examined how alignment influences looking patterns and speech onset times in a production and visual world eye-tracking study where choice of active or passive voice was forced. In actives, increased speech onset latency was selectively greater for object experiencer verbs than for subject experiencer verbs; while in passives, the pattern was reversed. An analysis of participants' looking patterns revealed more early looks to the subject in aligned configurations than in misaligned configurations. Taken together, Ferreira (1994) and Altmann & Kemper (2006) show consistent effects of alignment on structure choice, and Do & Kaiser (2019, 2022)'s results show that alignment influences on-line processes in production planning when choice of structure is constrained by the task context.

In sum, previous work on English has generally found evidence consistent with the effects of alignment in the domain of experiencer verbs in both processing and production. However, none of these studies to our knowledge has entirely disentangled the effect of an experiencer-first preference from the effect of an experiencer-subject preference, as all of these used simple monoclausal stimuli, or elicited simple monoclausal sentences. The exception is the reading time study discussed in Gennari & MacDonald (2008, 2009), which used SRCs and ORCs. However, they interpreted their results as reflecting the relatively greater accessibility of experiencer arguments given the specifics of their experimental design, which we would consider to be more generally compatible with the claims of experiencer-first approaches. This is largely because their study also manipulated the relative animacy of arguments, a factor which should interact strongly with any experiencer-related preference, since experiencers are animate by definition.

Work on languages other than English has found evidence more consistent with an experiencer-first preference. Studies on languages that allow simple OS clauses in active voice (i.e., active OS clauses where neither argument has been extracted) have shown that while SO orders are generally preferred, OS orders are relatively easier to process with object experiencer verbs compared to other verb types (Dutch: Lamers 2007 (sentence rating); German: Bornkessel et al. 2002 (ERP), Kretzschmar et al. 2012 (eye-tracking while reading); Italian: Dröge et al. 2014 (ERP); Spanish: Gat-

tei et al. 2015a (SPR, acceptability judgment), Gattei et al. 2015b (ERP), Gattei et al. 2017, 2022 (eye-tracking while reading); L1 Spanish: Mateu 2022 (picture-matching)). Such effects are consistent with a view that attributes the widely attested agent-first preference to general cognitive processes (Bickel et al. 2015; Bornkessel & Schlesewsky 2006; Bornkessel-Schlesewsky & Schlesewsky 2009; Hafri et al. 2013, 2018; Huber et al. 2023; Isasi-Isasmendi et al. 2023; Wilson et al. 2022), whereby it would naturally extend to experiencer arguments due to shared properties of agents and experiencers (including, e.g., obligatory animacy and saliency of mental states). Nevertheless, an experiencer-subject preference may yet be active even in languages that allow simple OS clauses. In particular, Verhoeven (2014) and Bader et al. (2017) found in production studies that even in German and Modern Greek (which allow simple OS clauses), the preferred way to realize an experiencer-first order was passivization, which also makes the experiencer the subject. Thus, there may be both experiencer-first and experiencer-subject preferences that apply during normal unconstrained production (and possibly, by extension, processing).

1.3 The present study

As mentioned above, researchers have looked at a range of factors relevant to evaluating the role of linking in comprehension, but none have carried out a within subjects design that targets all the relevant factors to disentangling an experiencer-first preference from an experiencer-subject preference in English. The main goal of our study is to build on this previous work by providing data that could do just that, with stimuli like the following.

(5)
a. That's the actor that the director particularly appreciates because of his dramatic emotional bearing. (active, subject experiencer)
b. That's the actor that the director particularly bothers because of his dramatic emotional bearing. (active, object experiencer)
c. That's the director that the actor is particularly appreciated by because of his dramatic emotional bearing. (passive, subject experiencer)
d. That's the director that the actor is particularly bothered by because of his dramatic emotional bearing. (passive, object experiencer)

In particular, we revisit English, using a new design that allows us to disentangle linear order alignment from grammatical function alignment. Our items, as in (5), used presentational object relative clauses (e.g., *That's the employee that the boss likes*), such that the first argument of the critical verb was the **object**, rather than the subject. This set-up leads to different predictions depending on one's views regarding the source of difficulty with misaligned configurations (e.g., an experiencer-first preference vs. an experiencer-subject preference). In particular, the two preferences we have considered would lead to exactly the opposite behavior in these configurations. In presentational ORCs, the experiencer-first preference would favor (5b) and (5c), since in these cases the first argument in the sentence bears the experiencer role. These sentences would thus be easier to process if there is a preference to link earlier arguments to a higher thematic role (experiencer) over a lower thematic role (theme). In contrast, (5a) and (5d) should be relatively harder, since the earlier argument is the theme, which is lower ranked than the later experiencer argument. This could lead to a need for reanalysis (Bornkessel & Schlesewsky 2006; Bornkessel-Schlesewsky & Schlesewsky 2009), since the earlier argument may be incrementally assigned a higher ranked thematic role, which in these sentences would need to be either de-assigned, or reassigned to the other argument so that the correct role could ultimately be assigned.

In contrast, an experiencer-subject preference would make exactly the opposite prediction. In this case, (5a) and (5d) should be easier to process, since these would link the higher ranked thematic role, experiencer, to subject position. In contrast, sentences like (5b) and (5c) would be harder to process, since these would show the opposite linking pattern. Regarding the on-line source of such an effect, we might speculate that building structures for surface-aligned configurations during parsing could be easier, since these configurations more transparently respect grammatical constraints. Representations that the parser constructs incrementally that do not respect such constraints would require revision before arguments could be integrated.

Our design has several other important features. We uniformly used object-headed ORCs with two definite, animate arguments, controlling for many important factors that interact with the processes of interest (cf. Gennari & MacDonald 2009; Wagers & Pendleton 2016). Furthermore, in this construction, the verb is the single word in all conditions at which participants get access to the necessary argument structure information to integrate both arguments with the verb's meaning. This gives us a single point in each sentence—the critical verb—which indexes the difficulty associated with the integration of argument structure information over the entire clause.

In addition, the fact that the head of the ORC was the main clause object meant that the head noun had the same grammatical function it had in the embedded clause in the main clause (in actives), or at least a more comparable one (in passives, because objects are closer to obliques on the grammatical function hierarchy than are subjects). Thus, there would not be conflicting preferences based on the grammatical function hierarchy's interaction with clause structure, as could arise with subject-headed ORCs, where the extracted noun would bear distinct grammatical functions in each clause.

Several studies have noted that object experiencer verbs are potentially ambiguous between eventive agent-theme readings and stative object experiencer readings (Belletti & Rizzi 1988; Bidgood et al. 2020; Darby 2016; Gattei et al. 2022; Landau 2010; Pesetsky 1995). Without explicitly controlling for this, participants might read object experiencer verbs as agent-theme verbs. It is thus important for the logic of our study that participants interpret object experiencer verbs as stative. Though we are aware of no way to force a stative reading of a potentially ambiguous sentence (Seth Cable, p.c.), we attempted to control this by using adverbs (e.g., *particularly*, *somewhat*, *greatly*, *really*, or *strongly*) that were all intended to be plausibly read as intensifiers. We considered these sufficiently intuitively biased toward stative readings to be of use. Another way we encouraged stative readings was by using the embedded verb/passive auxiliary in the simple present tense. Even if our controls were not successful, this would predict that object experiencer verbs should behave like agent-theme verbs—that is, they would show aligned configurations in active sentences and misaligned configurations in passive sentences. This would make predictions consistent with the experiencer-first preference, but go against the predictions of the experiencer-subject preference. As our results ultimately were more consistent with the experiencer-subject preference, and not at all indicative of an experiencer-first preference, we may put aside this potential confound.

2. Experiment 1

Experiment 1 aimed to distinguish the effects of a potential experiencer-first preference from those of a potential experiencer-subject preference. Specifically, an experiencer-subject preference predicts that in actives, object experiencer verbs should produce more difficulty than subject experiencer verbs; in passives, the opposite pattern should be found. In addition, modulo any main effect

of voice, we could predict subject experiencer verbs to produce more difficulty in passives than in actives, and the opposite for object experiencer verbs. Thus, an experiencer-subject preference might predict effects of voice nested within Verb type, as well as effects of Verb type nested within voice. An experiencer-first preference, on the other hand, would produce exactly the opposite pattern of results.

2.1 Materials

2.1.1 Stimuli

We constructed 32 items like those in (5) in a within-subjects 2 × 2 design, crossing Verb type (subject experiencer, object experiencer) and Sentence voice (active, passive). We used 32 verbs (16 subject experiencer, 16 object experiencer), based on Levin (1993)'s categorization[1]. A full list of items can be found in Appendix A.

(6) a. Subject experiencer verbs:
admire, *adore*, *appreciate*, *cherish*, *detest*, *dislike*, *distrust*, *dread*, *envy*, *esteem*, *fancy*, *idolize*, *loathe*, *respect*, *revere*, *tolerate*

b. Object experiencer verbs:
amaze, *annoy*, *appall*, *bore*, *bother*, *depress*, *disappoint*, *frighten*, *impress*, *interest*, *intrigue*, *irritate*, *perturb*, *repulse*, *terrify*, *unnerve*

Verbs were matched for length in characters across active and passive uses and verbal lemma frequency according to SUBTLEX part of speech counts (Brysbaert & New 2009) (length: $\mu_{\text{subjexp}} = 7.78, \mu_{\text{objexp}} = 8.28$; Welch's $t\,(60.78) = -1.36, p = 0.18$; frequency: $\mu_{\text{subjexp}} = 651.31, \mu_{\text{objexp}} = 640.44$; Welch's $t\,(29.96) = 0.03, p = 0.98$).

Across our 32 items, each pair of verbs occurred in 2 item sets. However, the items were counterbalanced to ensure that a participant saw only one token of each verb. Verbs occurred in different voices across lists; we thus ensured that length in characters did not differ between conditions across our lists as well (mean groups A and D: subjexp = 7.81, objexp = 8.31; Welch's $t\,(29.83) = -0.95, p = 0.35$; mean groups B and C: subjexp = 7.75, objexp = 8.25; Welch's $t\,(28.78) = -0.94, p = 0.36$).

Nouns were used in the same grammatical functions across subject experiencer and object experiencer sentences of the same voice, so that each argument occurred equally often as the theme and as the experiencer between participants.

Following the embedded clause verb was an adjunct phrase, headed by *because of*, *in spite of*, *due to*, *on account of*, or *despite*, intended for use as a spillover region. None of our effects of interest come from words following the spillover word.

Each sentence was paired with one of two comprehension questions, of the following form (using the questions for the example stimulus in (5)):

(7) a. Who appreciates someone? (active question, subject experiencer)
b. Who bothers someone? (active question, object experiencer)
c. Who is appreciated? (passive question, subject experiencer)
d. Who is bothered? (passive question, object experiencer)

[1] For subject experiencer verbs, we used Levin (1993)'s class 31.2 '*Admire* Verbs'; for object experiencer verbs, we used her class 31.1 '*Amuse* Verbs.'

(8) Answers: the director the actor (both orders)

Questions were constructed so that in order to answer them correctly consistently, participants would have to correctly link at least one argument of the embedded clause verb to the thematic role the embedded verb gave it. This allows us to measure successful comprehension of the linking configuration of each presentation of a stimulus. Question voice and the order of possible answers was counterbalanced within sentence conditions. Thus, while there were four conditions per sentence created by crossing Verb type with Sentence voice, there were 16 conditions per question, created by crossing those factors with Question voice (active, passive) and Answer order (correct first, correct second). We consider Question voice but not Answer order in our results.

We counterbalanced Question voice within items to account for the possibility of a same-voice advantage: if questions were uniformly in active (passive) voice, then subjects may have answered more accurately across the board in the active (passive) conditions. In addition, since questions used the same experiencer verbs as the sentences, alignment of either sort may affect processing of the questions as well, which could influence accuracy.

Items were split across 16 lists using a Latin Square design such that no participant saw each item in more than one condition. Participants saw pairs of subject experiencer and object experiencer verbs of similar lemma frequency counts in opposite conditions. For instance, *appreciate* (subject experiencer) occurred 4,861 times according to the SUBTLEX part of speech counts, and *bother* (object experiencer) occurred 5,102 times, making their raw frequency counts comparable. Thus, if a participant saw *appreciate* in the active sentence voice, passive question voice, correct answer on left condition; they would see *bother* in the passive sentence voice, active question voice, correct answer on right condition. Participants saw exactly one item containing each of the 32 critical verbs. We thus had 8 reading time measurements per condition per participant for the critical experimental sentences.

2.1.2 Fillers

Fifty-six unrelated fillers were included. Fillers were constructed such that participants would see a comparable number of total active and passive sentences and questions throughout the experiment. Answer order (left, right) was balanced within fillers. Five filler items had ambiguous questions for which either answer was considered correct. Altogether, participants read and responded to 88 items (32 experimental + 56 fillers).

2.2 Methods

2.2.1 Participants

140 undergraduate students from the University of Massachusetts Amherst participated in the experiment, receiving course credit. Eight participants' data was excluded due to a technical error. To qualify for analysis, participants had to be 18 years of age or older, L1 English speaking adults (16 exclusions), with normal or corrected to normal vision (10 exclusions) and no history of language disorders (6 exclusions). They also had to answer at least 7 out of 8 questions about the experiment's instructions correctly in a pre-experiment questionnaire (33 exclusions), which we decided prior to analysis. In addition, participants were excluded if they were enrolled in a course where the first author had discussed the design and purpose of the study (1 exclusion), or if they indicated that they had participated in a pilot using similar materials (4 exclusions). They also had to respond to

an open-ended post-experiment prompt designed to screen out attempted automated completion of the study to the satisfaction of the authors ("*Imagine you drove (or walked) from your house to the closest major shopping mall. Describe the most boring thing and the most interesting thing you would see along the way.*") (2 exclusions). Several participants were excluded on the basis of multiple criteria. Ultimately, 76 participants' data were analyzed (56 exclusions overall).

2.2.2 Procedure

Experiments were conducted online using the (now defunct) IbexFarm platform (Drummond 2011) using a centered word-by-word self-paced reading paradigm. Prior to each sentence, participants saw a fixation cross in the upper middle of the screen for 1000 ms, after which two dashes appeared in its place. Participants pressed the space bar when they were ready to begin the sentence, at which point the dashes would be replaced with the first word of the sentence; presentation was non-cumulative. Participants pressed the space bar to proceed to following words one at a time, until they reached the end of the sentence (indicated by a period). After each sentence, participants saw a comprehension question. Presented simultaneously with the question were two answers, one to the left and one to the right. Participants pressed the F key to choose the answer on the left or the J key to choose the answer on the right. Participants were given feedback on incorrect answers for filler items only, with the following message displayed for 1000 ms: "*Wrong answer. Please read slowly and carefully.*" For correct answers on fillers, participants saw a fixation cross for an additional 1000 ms instead of the warning message. No feedback was provided for experimental items.[2]

Upon arriving at the experimental website, participants consented, took a questionnaire about general demographic information and relevant exclusion criteria, and read the instructions, each page of which contained questions to ensure that they had read and understood that page. We used > 1 incorrect response to these questions as an exclusion criterion (see section 2.2.1).

Participants took two breaks during the experiment. They were told to take as much time as they needed for these, but to try not to take too long. After participants read and responded to all 88 items, they answered some open-ended questions about their impressions of the experiment, including one question designed to screen out automated completions.

2.3 Statistical analysis

We used deviation coding for our categorical predictors, as follows: Sentence/Question voice: active $= -0.5$, passive $= 0.5$; Verb type: subject experiencer $= -0.5$, object experiencer $= 0.5$. For nested comparisons, we set the values of the excluded conditions to 0.

For analysis, we consider log RTs for the critical verb and spillover word.[3] We also considered question accuracy. All RTs and question responses from eligible participants were analyzed.

We conducted two types of statistical analyses using R (R Core Team 2017), which produced essentially identical results. Here, we present the results of a Bayesian analysis using R's `brms` li-

[2] An anonymous reviewer asks to what extent giving feedback on only filler items may have drawn attention to the experimental items. In practice, accuracy for filler questions was very high overall (92.4%). Because of this, we believe participants would have been unlikely to notice a difference between fillers and experimental items in terms of feedback.

[3] Two items contained one argument noun phrase each that consisted of two words: "park ranger" and "rock star." We summed the reading times for these two words in the word-based measures to treat them as a single word. All of our critical effects came at least 2–3 words after these in the worst case, and none of our critical data came from RTs for these words. Given this, as well as the fact that this only affected 2 out of the 32 items, we do not think this compromises our results.

brary (Bürkner 2017, 2018, 2021); we also conducted a frequentist analysis using the R libraries `lme4` (Bates et al. 2015) and `lmerTest` (Kuznetsova et al. 2017), whose results we include in our supplementary materials on our OSF repository at osf.io/jq62t. For the Bayesian analysis, we used relatively uninformative priors, following Avetisyan et al. (2020). For the intercept, we used a prior of $\mathcal{N}(0, 10)$. For all predictors, interactions, and σ, we used priors of $\mathcal{N}(0, 1)$. For the correlation matrix prior, we used LKJ(2). For RTs, we fit multilevel MCMC models on the critical verb and the spillover word using a log-normal linking function. We report 95% Credible Intervals for each effect (with the exception of the main effects in the nested models, which were always essentially identical to the main effects in the corresponding crossed models, as expected). We interpret effects whose 95% CIs do not span 0.

2.4 Results

Prior to analysis, we planned to remove any participant whose overall question response accuracy (including fillers) was not reliably above chance (i.e., 50%). This was determined with a χ^2 test. We took a p-value of of less than 0.2 as reasonable evidence that they were not responding at chance. No participants were excluded by this criterion (low: 65.1% (54/83), high: 95.1% (79/83), μ = 67.8%, σ = 5.63%).

We present models containing the factors Sentence voice and Verb type (as well as Question voice, for response accuracy only), both crossed and nested. We considered both models nesting Sentence voice in Verb type, and Verb type in Sentence voice, since both ways of comparing alignment (across Verb types within a specific voice, and across a voice for a specific Verb type) are relevant for assessing the experiencer-subject preference. This is because the alignment preferences we consider make predictions for both nestings. For instance, an experiencer-subject preference predicts active subject experiencer sentences should be easier than active object experiencer sentences, and vice versa for passive sentences. The model that assesses this prediction nests Verb type within Sentence voice. A second prediction the experiencer-subject preference makes is that active subject experiencer sentences should be easier than passive subject experiencer sentences, and vice versa for object experiencer verbs. The model that assesses this prediction nests Sentence voice within Verb type. In each case, an experiencer-first preference would predict exactly the opposite preferences, given the nature of our items.

Figure 1 displays word-by-word log RTs, with the x-axis aligned at the critical verb.[4] In all plots, bars represent standard errors.

[4]Note that the word-by-word plot contains two words in both conditions that are not included in the example item used for the x-axis label text. This is because in the post-spillover region, the number of words was not equal across items. None of the effects we discuss come from any word after the spillover word.

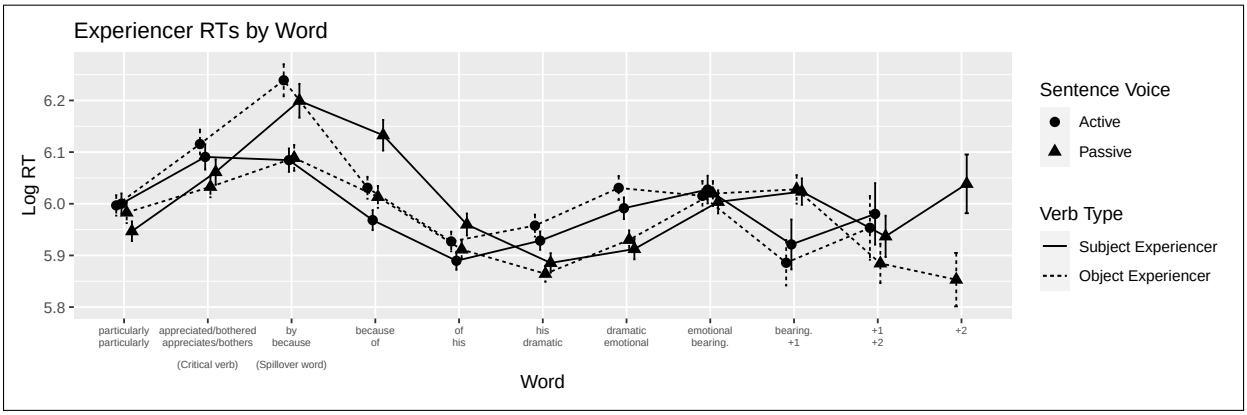


Figure 1: Log RTs by word, aligned at critical verb (Experiment 1)

Question response accuracy is shown in figure 2. While our analysis is concerned only with the relative difficulty of different conditions rather than absolute difficulty, we note that performance in the subject experiencer passive condition appears to be around chance, and performance in the object experiencer active condition is not much better. We speculate that participants found these conditions sufficiently difficult to comprehend (at least with regards to the thematic roles of the arguments) that they simply guessed most of the time.[5]

[5]Two anonymous reviewers ask why we did not exclude RTs from sentences with incorrect answers from analysis, given the low level of accuracy in some conditions. While we believe that there is good reason to include RTs from sentences with incorrect responses as reflecting failed attempts at comprehension, we conducted an analysis that omitted such RTs. The pattern of results was essentially identical to the analyses conducted without removing RTs for these sentences. As such, we present the results from the full dataset in the text.

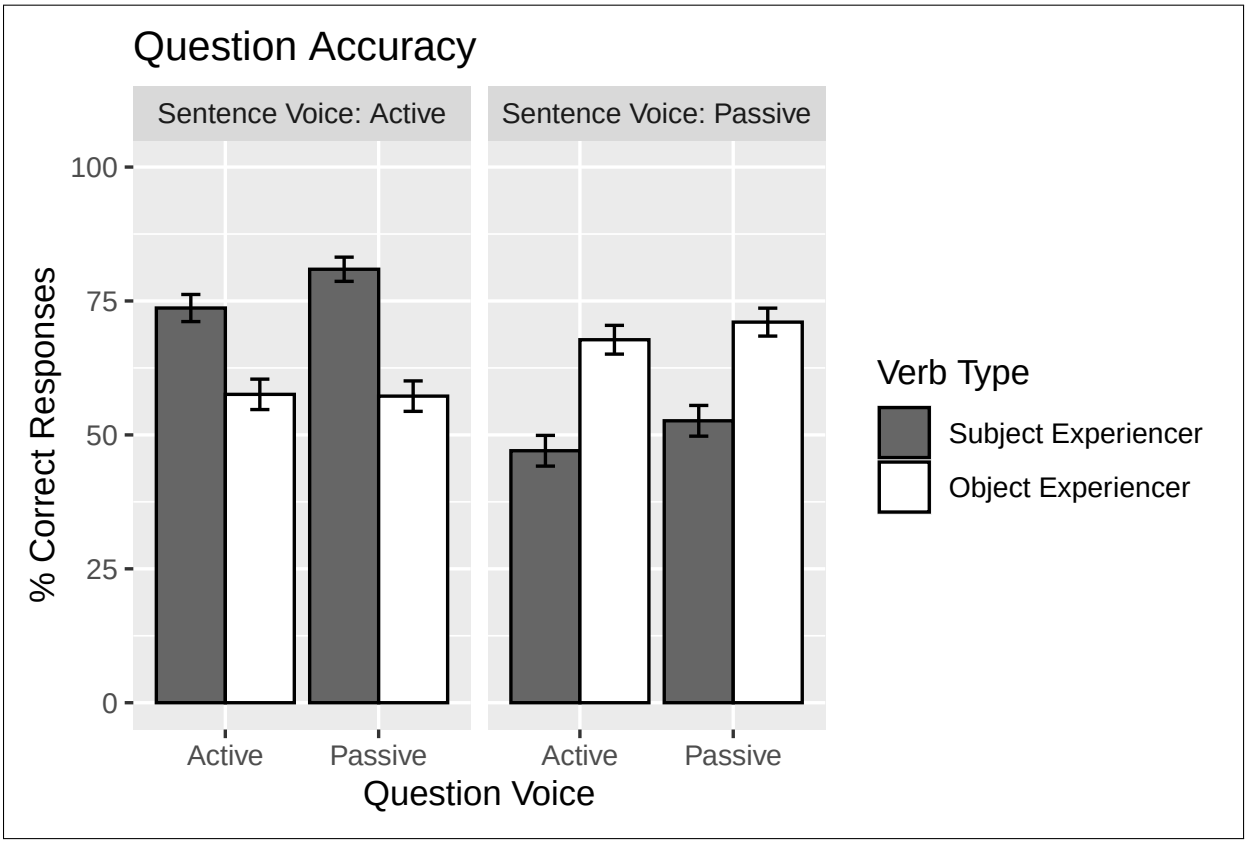


Figure 2: Question accuracy by Verb type, Sentence voice, and Question voice (Experiment 1)

2.4.1 Critical verb

At the critical verb, there were longer overall reading times in the active condition than in the passive condition (95% CI: $[-0.10, -0.01]$). There was no association between reading times and Verb type, nor any interaction 1 (all 95% CIs span 0). (In this and all subsequent tables, values are rounded to two decimal places for presentation; e.g., −0.00 means that a value is < 0, but rounds to 0.00. Rounding while indicating the direction from which the rounding occurred is intended to optimize for a combination of visual clarity in tables, while facilitating evaluation of whether a 95% CI actually spans 0.)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	6.08	0.05	5.98	6.17
Verb type	−0.00	0.02	−0.04	0.04
Sentence voice	−0.06	0.02	−0.10	−0.01
Verb type × Sentence voice	−0.05	0.04	−0.13	0.03

Table 1: Crossed model for the critical verb without frequency (Experiment 1)

As there was no interaction, we do not consider the nested models for this word.

2.4.2 *Spillover word*

At the spillover word, an interesting pattern emerged, as shown in table 2. There was no link between Verb type or Sentence voice and RTs. Nevertheless, there was a link between RTs and an interaction between these two factors (95% CI: $[-0.39, -0.14]$). We resolved this interaction by both Verb type and Sentence voice.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	6.15	0.05	6.05	6.26
Verb type	0.02	0.03	−0.03	0.07
Sentence voice	−0.02	0.03	−0.08	0.04
Verb type × Sentence voice	−0.27	0.06	−0.39	−0.14

Table 2: Crossed model for spillover word without frequency (Experiment 1)

The model with Verb type nested in Sentence voice (table 3) showed that Verb type was associated with differences in RTs in both active and passive sentences, but in opposite directions.[6] This reflected the fact that in active voice sentences, subject experiencer verbs led to shorter RTs than did object experiencer verbs; while in contrast, in passive sentences, object experiencer verbs led to shorter RTs than did subject experiencer verbs.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Verb type (active)	0.15	0.04	0.08	0.23
Verb type (passive)	−0.11	0.04	−0.19	−0.03

Table 3: Verb type nested in Sentence voice model for spillover word without frequency (Experiment 1)

The model with Sentence voice nested within Verb type (table 4) showed that Sentence voice was associated with differences in RTs in sentences with subject experiencer verbs and in sentences with object experiencer verbs, but in opposite directions. In sentences with subject experiencer verbs, participants read the spillover word more quickly in active sentences than in passive sentences; while in sentences with object experiencer verbs, they read the spillover word more quickly in passive sentences than in active sentences.

[6] In this and future tables for nested models, we omit the intercept and main effects for simplicity. Though the estimates sometimes differed very slightly from model to model because of the random nature of MCMC sampling, estimates of the intercept and main effects remained very consistent compared across the fully crossed and nested models, making re-presenting them in each table largely redundant. Our full set of results, which include all estimates for all models, are included in our supplementary material.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Sentence voice (subjexp)	0.11	0.05	0.02	0.20
Sentence voice (objexp)	−0.15	0.04	−0.22	−0.08

Table 4: Sentence voice nested in Verb type model for spillover word without frequency (Experiment 1)

2.4.3 Question responses

For question accuracy, we fit a logistic multilevel model. In addition to the interaction between Sentence voice and Verb type, this model also included an interaction between this interaction and Question voice. There was an association of accuracy with Sentence voice (95% CI: $[-0.62, -0.18]$), reflecting overall fewer accurate responses to questions about passive voice sentences. Finally, accuracy was associated with an interaction between Verb type and Sentence voice (95% CI: $[1.49, 2.31]$).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	0.66	0.10	0.47	0.86
Verb type	−0.04	0.13	−0.31	0.22
Sentence voice	−0.40	0.11	−0.62	−0.18
Question voice	0.21	0.11	−0.01	0.42
Verb type × Sentence voice	1.90	0.21	1.49	2.31
Verb type × Question voice	−0.24	0.24	−0.70	0.24
Sentence voice × Question voice	0.01	0.23	−0.45	0.48
Verb type × Sentence voice × Question voice	0.35	0.38	−0.39	1.11

Table 5: Crossed model for question accuracy without frequency (Experiment 1)

We resolved the interaction of Verb type and Sentence voice by considering models nesting Verb type within Sentence voice and nesting Sentence voice within Verb type (which included interactions with Question voice).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Verb type (active)	−1.02	0.17	−1.36	−0.67
Verb type (passive)	0.93	0.16	0.62	1.26
Verb type (active) × Question voice	−0.42	0.30	−1.02	0.17
Verb type (passive) × Question voice	−0.04	0.29	−0.61	0.54

Table 6: Verb type nested in Sentence voice model for question accuracy without frequency (Experiment 1)

Accuracy was associated with Verb type in both active and passive voice sentences, but in opposite directions. This reflected that in active sentences, participants responded correctly more often

to subject experiencer verbs than to object experiencer verbs (95% CI: $[-1.36, -0.67]$); in passive sentences, participants showed the opposite pattern, responding correctly more often to object experiencer verbs than to subject experiencer verbs (95% CI: $[0.62, 1.26]$). There was no interaction of these effects with Question voice.

Considering the model that nested Sentence voice within Verb type next (table 7), we found voice was associated with accuracy for both subject experiencer and object experiencer verbs, but in opposite directions. This reflected the fact that participants responded accurately more often to active sentences containing subject experiencer verbs than to active sentences containing object experiencer verbs (95% CI: $[-1.68, -1.05]$), while they responded accurately more often to passive sentences containing object experiencer verbs than to passive sentences containing subject experiencer verbs (95% CI: $[0.29, 0.88]$). There was no interaction between either of these effects and Question voice.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Sentence voice (subjexp)	−1.37	0.16	−1.68	−1.05
Sentence voice (objexp)	0.59	0.15	0.29	0.88
Sentence voice (subjexp) × Question voice	−0.18	0.27	−0.71	0.36
Sentence voice (objexp) × Question voice	0.22	0.34	−0.45	0.89

Table 7: Sentence voice nested in Verb type model for question accuracy without frequency (Experiment 1)

2.5 Discussion

Experiment 1 results showed that active subject experiencer verbs generally led to faster RTs and better comprehension than active object experiencer verbs and passive subject experiencer verbs. Conversely, active object experiencer verbs generally led to faster RTs and better comprehension than passive subject experiencer verbs and active object experiencer verbs.

To account for possible effects of frequency (cf. Gennari & MacDonald 2009; Gibson et al. 2019; Hale 2001; Levy 2008a; MacDonald 2013), we conducted a corpus study to obtain verb frequency by sentence voice (detailed in Appendix B). While our items controlled for lemma frequency of subject and object experiencer verbs, they did not control for the frequency with which those verbs occurred in a particular voice. An alternative explanation for our results could have been due to object experiencer verbs occurring more frequently in passive than in active voice, and vice versa for subject experiencer verbs (cf. Ferreira 1994), if we assume that more frequently encountered configurations are easier to process (e.g., Gennari & MacDonald 2009; Levy 2008a; MacDonald 2013). Nevertheless, including frequency in our models did not alter the pattern of results, arguing against this being a full explanation of our data. Detailed results are presented in Appendix C.

For spillover word RTs and question accuracy, we found a crossover interaction, with all resolutions of the interaction between Sentence voice and Verb type showing effects whose CIs did not overlap with 0. The directionality of these effects is consistent with an experiencer-subject preference, and inconsistent with an experiencer-first preference. The prior work we discussed sometimes found effects consistent with ours (e.g., Cupples 2002; Do & Kaiser 2019, 2022; Ferreira 1994; Gennari & MacDonald 2008, 2009; Grodzinsky 1995; Grodzinsky et al. 1991), and sometimes found

effects inconsistent with ours (e.g., Carrithers 1989), or effects more consistent with an experiencer-first preference (e.g., Bader et al. 2017; Bornkessel et al. 2004; Bornkessel & Schlesewsky 2006; Bornkessel et al. 2002, 2003, 2005; Bornkessel-Schlesewsky & Schlesewsky 2009; Ferreira 2003; Lamers 2007; Mateu 2022; Verhoeven 2014). A likely explanation for these differences is that previous studies did not always carry out a full within-subjects design crossing Verb type and Sentence voice, did not consistently control for effects of animacy and definiteness, and did not have a design that disentangled the effects of an experiencer-first preference from the effects of an experiencer-subject preference. Our study fills this gap, and shows that an experiencer-subject preference is active in English. To the extent that there may an independent experiencer-first preference, we have shown that in our study, it is entirely subsumed by the experiencer-subject preference.

In particular, in our data, participants processed sentences more easily when the experiencer came *second*, as in ORCs with active subject experiencer verbs and passive object experiencer verbs. In contrast, the conditions where the experiencer occurred first were relatively harder. This result is thus more consistent with an experiencer-subject preference than with an experiencer-first preference.[7] This lends some support to our interpretation of Verhoeven (2014)'s and Bader et al. (2017)'s production studies, who found that participants' preferred strategy for producing an experiencer-first sentence was passivization even in German and Modern Greek, which permit simple OS clauses. In addition, factors such as saliency, animacy, and accessibility that may contribute to the experiencer-first preference were controlled for in our design, meaning that participants would not have reason to display a preference that wasn't specifically targeted at the argument bearing the experiencer thematic role coming first, as opposed to an animate argument coming first. If the experiencer-first preference is not about experiencer arguments per se, and instead about the fact that experiencer arguments are necessarily animate while theme arguments are not, that could explain why our data do not show it, since our items used two animate nouns as arguments of the critical verb.

Nevertheless, we recall that evidence for an experiencer-first preference that disentangles it from an experiencer-subject preference has been found in comprehension studies of non-English languages, including Dutch, German, Italian, and Spanish (Bornkessel et al. 2002, 2003, 2005; Dröge et al. 2014; Gattei et al. 2022, 2015a, 2017, 2015b; Kretzschmar et al. 2012; Lamers 2007; Mateu 2022). Interestingly, Bornkessel & Schlesewsky (2006) note that in German, they found an experiencer-first effect only with overt morphological case marking, which English (mostly) lacks (see also Bader & Bayer 2006). We speculate that the presence of certain cues in a language (like overt morphological case) may lead speakers to develop preferences conditioned on those cues. English speakers might thus develop preferences based more on grammatical functions, which are more easily detectable in English without the aid of case marking, due to its more rigid word order compared to German. Determining how a language's grammatical properties influence its speakers' on-line processing strategies would be a fruitful way of carrying forward this line of research.

Unlike our results, Gattei et al. (2022) found no effects of alignment on comprehension accuracy (although they did in previous studies (Gattei et al. 2015a, 2017, 2015b)), despite finding effects on late eye-tracking measures. They propose this is because the comprehension questions in their earlier studies mentioned both verbal arguments, while the comprehension questions in Gattei et al. (2022) mentioned only one argument. They posit that the lower accuracy found in the earlier studies

[7] Our data are similarly incompatible with a proposal that comprehenders should find sentences easier when affected arguments are extracted, as suggested to us by Rebecca Tollan (p.c.). In particular, our participants found processing and comprehension easier when the unaffected *theme* was extracted, compared to when the affected experiencer was extracted.

was due to participants needing to process linking information for all arguments in both sentences and questions, leading to greater difficulty than when only one argument was mentioned. However, our study's questions also mentioned only one participant, and yet we found clear effects of alignment. In addition, no interaction including Question voice and Verb type was found, suggesting that reprocessing linking in our questions is unlikely to have been responsible for the effects we observe. Instead, it seems likely that for our study, difficulty induced by the misaligned sentences themselves led to greater comprehension difficulty, producing lower accuracy in these conditions. This supports a view that alignment affects not only on-line processing (consistent with Gattei et al.'s findings), but also comprehension processes, to the extent that it might lead to comprehension failure.

3. Experiment 2

The results of experiment 1 provide evidence for the importance of an experiencer-subject preference in comprehension processes and on-line processing. Nevertheless, such results were not consistent with all previous comprehension studies, which sometimes found linear order effects or found effects for Verb type only for passive voice. Thus, it is important to show that our results were not spurious. For this reason, we conducted an additional study to examine whether we would find similar effects with a speeded grammaticality judgment measure. This measure is somewhat in-between an on-line measure like reading time and an off-line measure like question responses (which were not explicitly time-limited in our first experiment).

Importantly, judging whether a sentence is acceptable arguably presents a simpler task than attempting to assign it a sensible meaning. This distinction in tasks is relevant for interpreting the results of experiment 1: Bader & Meng 2018 and Meng & Bader 2021 showed that different experimental tasks can reveal divergent patterns of comprehension difficulty on misaligned agent-patient passive sentences (cf. Ferreira 2003). In their studies, an explicit comprehension question task suggested that comprehenders had difficulty in mapping the agent and patient roles to the arguments of a passive sentence. However, comprehenders experienced much less (or no) such difficulty when asked to perform a plausibility judgment task on the basis of that same mapping. This set of findings raises the possibility that the comprehension task in experiment 1 may have exaggerated the difficulty associated with processing misaligned configurations, or reflects a later stage of the comprehension process distinct from early processing of the marked configuration. On the other hand, judging whether a sentence is grammatical does not, strictly speaking require explicitly reporting who did what to whom, since a sentence can be judged as grammatical based entirely on its form. We thus might expect participants to show less difficulty with marked configurations in a grammaticality judgment task, since fully successful comprehension is not required to judge a sentence's acceptability (see Ferreira & Patson 2007).

In experiment 2, we critically test the difficulty associated with marked configurations by investigating how they influence behavior in a task that does not require correctly assigning the arguments to the thematic roles a verb provides. For this reason, seeing no link between perceived acceptability and structure/verb type in experiment 2 could indicate that participants' performance in experiment 1 was due to the additional processing involved in answering comprehension questions in certain configurations. On the other hand, seeing a link between acceptability and structure/verb type in experiment 2 would indicate that the experiencer-subject preference is active even in contexts where achieving accurate linking is not required by the task context, indicating that the difficulty associated with misaligned configurations generalizes to arguably different task contexts.

3.1 Materials and methods

3.1.1 Stimuli

We reused the 32 items from experiment 1 for experiment 2. The only difference was that there were no comprehension questions. This meant that each item occurred in 4 conditions, which were the result of crossing Verb type (subject experiencer, object experiencer) and Sentence voice (active, passive).

3.1.2 Fillers

114 unrelated fillers were included, some of which were from unrelated experiments. 56 of the fillers were ungrammatical, with the remainder being grammatical. Thus, participants saw a total of 146 items.

3.1.3 Participants

64 participants were recruited from Prolific (prolific.co). We used Prolific's built-in prescreeners to control who was able to view the study. To see the study, potential participants had to be native-born United States citizens whose first language was English. They had to be fluent in English, and have no history of language-related disorders or reported literacy difficulties, and to not have participated in a previous pilot study using the same items. Each participant was paid 5 USD in compensation, based on an average completion time of around 30 minutes.

3.1.4 Procedure

We hosted our experiment using PCIbex (farm.pcibex.net) (Zehr & Schwarz 2018). Eligible participants who decided to participate would follow a link to our study page. Upon arrival, they would consent, and then read through a set of instructions, which contained a 6-item questionnaire about the instructions designed to ensure participants were reading and understanding them. Participants were told to rate sentences for grammaticality, defined as sounding like natural English they could imagine someone might use. They were told they would have to respond within 2 seconds following the end of each sentence, and that they should respond according to their gut feeling about a sentence's grammaticality, rather than according to prescriptive norms. Participants used the F key to indicate that a sentence sounded grammatical, and the J key to indicate that it sounded ungrammatical.

Prior to beginning a trial, participants saw a fixation cross in the upper middle of the screen for 1000 ms. After this delay, the sentence was shown using rapid serial visual presentation in the center of the screen at a rate of 325 ms per word. Following the sentence, participants saw the question *Was the sentence grammatical?*, which they responded to using one of the two answer keys. Participants were given 2000 ms to respond, after which they were unable to rate the sentence, and would see a message for 2000 ms that they had timed out and should judge sentences more quickly.

After they had rated all 146 items, participants were given a chance to respond to open-ended follow up questions, including an open-ended prompt designed to screen out attempted automated completions (this used the same prompt as in experiment 1). In all, the experiment took about 30 minutes.

3.2 *Results*

To be included in the analysis participants had to answer at least 5 out of 6 questions from the pre-experiment questionnaire correctly (0 exclusions), and answer the question designed to screen out automated completions to the satisfaction of the authors (0 exclusions). Thus, data from all 64 participants were included in the analysis.

Not every sentence was rated prior to the time-out limit, which meant that 22 observations (0.01%) were missing. The missing observations were distributed roughly equally across the 4 conditions: 6 subjexp active, 6 subjexp passive, 4 objexp active, and 6 objexp passive.

Our dependent measure was participants' grammaticality judgments. To examine the responses, we fit multilevel logistic regressions using the `brms` package, with Response as our dependent variable ("Grammatical" and "Ungrammatical," coded as 1 and 0). We used deviation coding for our categorical predictors, as follows: Sentence voice: active = 0.5, passive = −0.5; Verb type: subject experiencer = 0.5, object experiencer = −0.5. For nested comparisons, we set the values of the conditions to be excluded to 0. We used the same priors we had used in experiment 1. We used the maximal random effects structure for each model. We also conducted frequentist analyses using R's `lme4` and `lmerTest` packages, which were consistent with the Bayesian analyses; these are included in our supplementary materials. As for experiment 1, we report estimates, estimated error, and 95% CIs for each effect. We interpret effects whose 95% CIs do not span 0.

Figure 3 shows the percent of "grammatical" responses by condition.

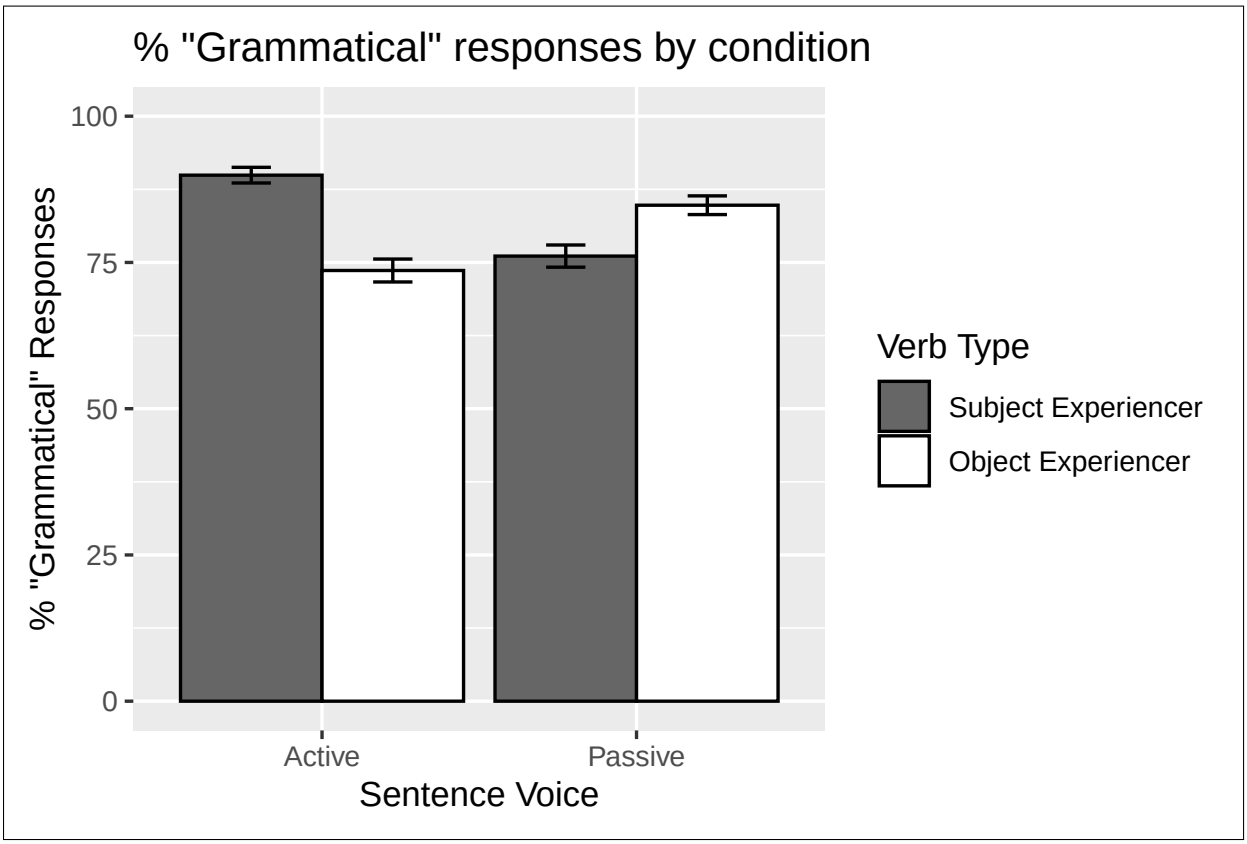


Figure 3: Percent "grammatical" responses by condition (Experiment 2)

Table 8 shows the results of the model crossing Sentence voice and Verb type.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	2.04	0.20	1.67	2.44
Verb type	−0.38	0.18	−0.75	−0.02
Sentence voice	0.17	0.19	−0.21	0.53
Verb type × Sentence voice	−1.86	0.40	−2.65	−1.08

Table 8: Crossed model for Response without frequency (Experiment 2)

Grammaticality judgments were associated with Verb type (95% CI: $[-0.75, -0.02]$). This reflected the fact that participants rated as grammatical a greater proportion of sentences containing subject experiencer verbs compared to sentences containing object experiencer verbs (subjexp: $\mu = 0.830$, $\sigma = 0.0118$; objexp: $\mu = 0.792$, $\sigma = 0.0128$). This main effect was qualified by an interaction between Verb type and Sentence voice (95% CI: $[-2.65, -1.08]$).

To examine this interaction, we fit models nesting Verb type in Sentence voice and nesting Sentence voice in Verb type.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Verb type (active)	−1.44	0.32	−2.08	−0.83
Verb type (passive)	0.62	0.24	0.14	1.09

Table 9: Verb type nested in Sentence voice model for Response without frequency (Experiment 2)

This model nesting Verb type in Sentence voice showed an association between judgments and Verb type in both active (95% CI: $[-2.08, -0.83]$) and passive sentences (95% CI: $[0.14, 1.09]$), but in opposite directions. This reflected the fact that in active sentences, participants responded that sentences were grammatical more often when they contained an object experiencer verb than when they contained a subject experiencer verb, but that the pattern was the opposite in passive sentences.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Sentence voice (subjexp)	1.16	0.29	0.60	1.73
Sentence voice (objexp)	−0.86	0.28	−1.42	−0.33

Table 10: Sentence voice nested in Verb type model for Response without frequency (Experiment 2)

The model nesting Sentence voice in Verb type similarly showed an association between judgments and Sentence voice with both subject experiencer (95% CI: $[0.60, 1.73]$) and object experiencer verbs (95% CI: $[-1.42, -0.33]$), though in opposite directions. This reflected the fact that participants judged sentences containing subject experiencer verbs as grammatical more often when the verb occurred in the active voice than when it occurred in the passive voice. For object experiencer verbs, the pattern was flipped, with participants judging sentences containing object experiencer

verbs as grammatical more often when the verb occurred in the passive voice than when it occurred in the active voice.

Overall, this pattern of results matches those we found from the models using only categorical predictors in experiment 1, particularly the models for the spillover word RTs and question accuracy, which found a similar crossover interaction (though without any main effects). We return to this point in our discussion.

As in experiment 1, we used results from our corpus study (see Appendix C to fit models that included frequency as a predictor. Like for experiment 1, this did not affect the pattern of results.

3.3 Discussion

Experiment 2 examined how English speakers judged the grammaticality of sentences with experiencer verbs under time pressure, and how their judgments depended on Verb type and Sentence voice, in a context where achieving a veridical representation of the meaning of the sentence was irrelevant to accuracy.

We found links between judgments and Sentence voice and Verb type in line with the experiencer-subject preference. Participants were more likely to judge sentences where the experiencer was the subject of the critical verb as grammatical compared to sentences where it was not, in both active and passive structures. Notably, these are precisely the configurations where in our items, the experiencer occurs second. These results, then, like the results of experiment 1, are not consistent with an experiencer-first preference.

The pattern of results was consistent with what we found in experiment 1. This was despite the fact that experiment 1's task (answering comprehension questions probing thematic role assignments) would seem to encourage deeper processing of thematic role assignments than experiment 2's task (judging whether a sentence is grammatical). Thus, even when controlling for the particular kind of difficulty that explicit comprehension questions targeting thematic role assignment might raise (Bader & Meng 2018; Meng & Bader 2021), we find that participants preferred experiencer-subject configurations over experiencer-first configurations. We take these results to indicate that the experiencer-subject preference found in experiment 1 is not related to any particular difficulty prompted by the comprehension question task in experiment 1, or to any late stage comprehension processes that are specific to that task.

4. General discussion

We conducted one self-paced reading experiment and one speeded grammaticality judgment experiment to deconfound an experiencer-first preference from an experiencer-subject preference. In both our experiments we found evidence in support of an experiencer-subject preference: configurations that aligned thematic roles with grammatical functions (e.g., passive object experiencer verbs, active subject experiencer verbs) were easier to process, more accurately comprehended, and rated more acceptable than were configurations where these were misaligned (e.g., active object experiencer verbs, passive subject experiencer verbs). Crucially, an experiencer-first preference would have been expected to lead to us observing exactly the opposite pattern of results.

Our results thus support an experiencer-subject preference more directly than previous work. No prior study we are aware of has controlled for animacy, definiteness, and the difference between experiencer-first and experiencer-subject preferences in a within-subjects design; our results fill this gap in the prior literature, and thus provide clear support for an experiencer-subject preference in

processing. Nevertheless, our results are not inconsistent with the substantial body of work showing an experiencer-first preference for languages other than English, nor indeed with such a preference existing for English; alignment occurs on many hierarchies simultaneously, which may interact in interesting ways with the thematic role and grammatical function hierarchies we focused on. However, our results have shown that if there is an experiencer-first preference in English, it must be quite weak compared to the experiencer-subject preference, which appears robust in our results.

Our results raise the question of how exactly thematic roles' alignment with grammatical functions influences comprehension processes. In particular, it is worth interpreting the experiencer-subject preference in light of the prior literature we reviewed. If we now understand these results to support an experiencer-subject preference, a fairly consistent picture emerges: sentences that exhibit alignment between thematic roles and grammatical functions seem to facilitate comprehension and production alike, are acquired more readily, and are less susceptible to impairment in aphasic individuals. Given this, we suggest the general picture that arises is most consistent with a framework where grammatical and non-linguistic cognitive constraints operate in tandem to influence both production and comprehension directly. However, grammatical constraints cannot necessarily be reduced to these more general, non-linguistic constraints that relate to factors such as salience and accessibility of words. Instead, the grammatical constraints, for grammar-internal reasons, treat certain structures as more or less marked, and this directly influences processing. This general approach would propose a direct role for grammatical constraints in comprehension processes, and is consistent with recent constraint-based proposals (notably Wagers & Pendleton 2016 and Wagers et al. 2018), as well as the view that the grammar is deployed in on-line processing and comprehension (e.g., Momma & Phillips 2018).

According to this approach, violable constraints encode alignment across different hierarchies. While we have focused on how this idea has been investigated for experiencer verbs, the notion of alignment is far older than this application of it. Various sorts of grammatical hierarchies have been proposed and the importance of their alignment has been well investigated in psycholinguistic, generative, and descriptive studies (see, e.g., Aissen 1999; Christianson 2001; Christianson & Cho 2009; Christianson & Ferreira 2005; Hammerly 2020; Hammerly et al. 2022). There has been debate over whether the source of such hierarchies resides in the grammar itself or in the interaction of non-linguistic cognitive constraints with the grammar (e.g., Aissen 1999; Minkoff 1994 *et seq.*). Our results seem to be most compatible with a view where the relevant hierarchies and associated constraints are primarily grammatical objects, given that our design controlled for other potential confounding factors (such as an experiencer-first preference, which could arise due to non-linguistically-specific constraints). In particular, while it may be possible to derive a thematic role hierarchy on language-external conceptual grounds (as suggested to use by an anonymous reviewer) and it may also be possible to derive a preference to put animate arguments earlier based on general properties of memory and salience that are not specific to language, it seems less plausible that the grammatical function hierarchy could be derived in this way.

In other words, the view most compatible with our results is that the preference for alignment between grammatical functions and thematic roles does not come from factors related to ease of retrieval, such as animacy, saliency, and event representations. Instead, the preference is more likely to come from argument linking principles like those proposed in Baker (1988) and Perlmutter & Postal (1984). These argument linking principles are not easily reduced to non-linguistic cognitive constraints, because they make reference to a grammatical function hierarchy that is hard to derive from non-linguistic cognitive principles.

There are several caveats worth noting. First, we are unaware of any production study that has disentangled the predictions of the experiencer-first preference from those of the experiencer-subject preference. This would be important for establishing a role for thematic role alignment with grammatical functions in production, as opposed to an experiencer-first bias. We could imagine that the unavoidable differences between production and comprehension could make one preference more active in the production modality, since starting from the task of encoding a message linguistically could lead to a larger impact of factors related to salience and accessibility of concepts associated with words. Second, the work here is based on English, and extending the present design to languages other than English is critical. Though there is, fortunately, much work on the role of alignment—even thematic role-grammatical function alignment—in non-English languages (Bader et al. 2017; Bornkessel et al. 2002; Christianson 2001; Christianson & Cho 2009; Christianson & Ferreira 2005; Dröge et al. 2014; Gattei et al. 2022, 2015a, 2017, 2015b; Hammerly 2020; Hammerly et al. 2022; Isasi-Isasmendi et al. 2023; Kretzschmar et al. 2012; Lamers 2007; Mateu 2022; Temme & Verhoeven 2016; Verhoeven 2010, 2014, 2015), such work may also be subject to the same sorts of concerns that motivated our study regarding the possibility of alternative explanations such as an experiencer-first preference. Thus, it remains to be seen when the preferences we find evidence for in our experiments extend to other languages and when they do not. While we discussed previous work that has found evidence for an experiencer-first preference that goes beyond an experiencer-subject preference in non-English languages, it remains an open question what the relative strength of these preferences is. If speakers of these languages are like speakers of English, then the experiencer-first preference would be predicted to be fairly weak compared to the experiencer-subject preference. On the other hand, we might expect that language specific factors could tune the relative strength of these preferences in interesting ways.

A final contribution of our design was the use of certain techniques to try and induce a stative reading of our object experiencer verbs, following work that treats object experiencer verbs as ambiguous between being true experiencer verbs (in stative readings) and agent-theme verbs (in eventive readings) (Belletti & Rizzi 1988; Dowty 1991; Landau 2010; Pesetsky 1995). Our pattern of results is consistent with our attempt to induce stative readings being successful. We believe it would be interesting for future work to compare processing and comprehension of object experiencer verbs under different kinds of readings to see if there is psycholinguistic evidence for such a split (cf. Bidgood et al. 2020; Darby 2016). Indeed, Gattei et al. (2022) report evidence consistent with the view that eventive (i.e., agent-theme) uses of object experiencer verbs display different alignment behavior compared to stative uses in online processing in Spanish, as induced by a within-verb manipulation of case assignment (accusative vs. dative case on the experiencer argument). Extending this approach to other languages could determine if this pattern is specifically linked to overt case marking or is more general.

An anonymous reviewer raises an alternative account as possibly consistent with our results: in particular, suppose that what causes difficulty is not alignment per se, but movement of an argument bearing a higher thematic role compared to an argument bearing a lower thematic role. Our results do not categorically rule this out, but we find it unlikely. In particular, the well-known advantage of subject relative clauses compared to object relative clauses has been established in a variety of studies that happen to examine agent-patient verbs (e.g., Staub 2010; Wagers & Pendleton 2016; Wagers et al. 2018). If participants found it more difficult to move arguments with higher ranked thematic roles, we might have expected the opposite pattern, with ORCs showing an advantage. Of course, there is a clear potential confound in this case—clause structure—which we acknowledge. More

generally, though, we note that the discourse effect of extraction is to foreground the extracted argument, and to background the clause it is extracted from. This alternative explanation, then, would amount to saying that at a discourse level, comprehenders prefer to interpret sentences where less prominent arguments are foregrounded and more prominent arguments are backgrounded, which seems intuitively unlikely to us. However, while we believe this alternative to be unlikely for these reasons, we admit that our data do not rule it out.

5. Conclusion

We conducted two experiments examining how the linking properties of experiencer verbs influence sentence processing, comprehension, and grammaticality judgments. All measures displayed consistent results: sentences with object experiencer verbs in active voice were harder to process and understand, and considered grammatical less often, than sentences with object experiencer verbs in passive voice and sentences with subject experiencer verbs in active voice. In addition, sentences with subject experiencer verbs in *passive* voice were harder to process and understand, and considered grammatical less often, than sentences with subject experiencer verbs in active voice and sentences with object experiencer verbs in passive voice. These results are consistent with an experiencer-subject preference, and show no experiencer-first preference in the processing of English, thereby providing empirical evidence that discriminates between two competing claims regarding the source of such effects in English. Previous designs for English experiments confounded these two preferences, though deconfounded evidence from other languages, such as German, Dutch, and Spanish, supports an experiencer-first preference. This raises the question of *why* preferences such as these may differ cross-linguistically, which we leave for future research. Interestingly, evidence has been found that the agent-first preference is cross-linguistically robust, even in languages where agents are relatively less likely to occur in the first argument position (Bickel et al. 2015; Huber et al. 2023; Isasi-Isasmendi et al. 2023). Why the experiencer-first preference should differ in this regard deserves investigation.

Our results are consistent with the view that the preference to align thematic roles with grammatical functions comes from language-specific argument linking constraints (e.g., Baker 1988; Perlmutter & Postal 1984). However, our results are compatible a number of theoretical frameworks that recognize a role for alignment in linking processes. While we motivated our study with a discussion of Belletti & Rizzi (1988)'s influential work in the Government and Binding Theory framework, our results do not crucially rely upon this particular implementation. They are compatible with any approach to linking that recognizes a thematic role hierarchy that ranks experiencers above themes, a grammatical function/syntactic hierarchy that ranks subjects above objects and obliques, and an influence of these hierarchies on processing, whether these ideas be expressed in neo-constructionist, lexicalist, construction grammar, or some other approach to argument structure. What is crucial is that such constraints be recognized not as merely describing possible linking patterns in canonical (active) sentences, but as affecting how participants resolve argument structure dependencies in real-time in canonical (active) and non-canonical (passive) sentences alike. The general picture that arises is consistent with views that recognize grammatical constraints as playing an active role in real-time language processing and comprehension (e.g., Lewis & Phillips 2015; Momma & Phillips 2018; Phillips 1996).

References

Aissen, J. (1999). Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory*, *17*, 673–711.

Altmann, L. J. P., Mullin, D. A., & Mann, T. (2004). Minimal effects of order of noun activation on sentence production. Poster presented at the Psychonomic Society, Minneapolis, MN.

Altmann, L. P. J., & Kemper, S. (2006). Effects of age, animacy and activation order on sentence production. *Language and Cognitive Processes*, *21*(1/2/3), 322–354.

Ambridge, B., Bidgood, A., & Thomas, K. (2021). Disentangling syntactic, semantic and pragmatic impairments in ASD: Elicited production of passives. *Journal of Child Language*, *48*, 184–201.

Argaman, V., & Pearlmutter, N. J. (2002). Lexical semantics as a basis for argument structure frequency biases. In P. Merlo, & S. Stevenson (Eds.) *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*, vol. 4 of *Natural Language Processing*, (pp. 303–324). Amsterdam/Philadelphia: John Benjamins.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, 1–18.

Bader, M., & Bayer, J. (2006). *Case and Linking in Language Comprehension: Evidence from German*, vol. 34 of *Studies in Theoretical Psycholinguistics*. Dordrecht: Springer.

Bader, M., Ellsiepen, E., Vasiliki, K., & Portele, Y. (2017). Filling the prefield: Findings and challenges. In C. Freitag, O. Bott, & F. Schlotterbeck (Eds.) *Two Perspectives on V2: The Invited Talks of the Deutsche Gesellschaft für Sprachwissenschaft 2016 Workshop "V2 in Grammar and Processing: Its Causes and Its Consequences"*, (pp. 27–49). Konstanz: University of Konstanz.

Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1286.

Baker, M. (1988). *Incorporation: A Theory of Grammatical Function Changing*. Chicago, IL: The University of Chicago Press.

——— (1989). Object sharing and projection in serial verb constructions. *Linguistic Inquiry*, *20*(4), 513–553.

Balogh, J., & Grodzinsky, Y. (1996). Varieties of passives in agrammatic Broca's aphasia: ϑ-grids, arguments, and referentiality. *Brain and Language*, *55*(1), 54–56. In *Poster Session 1: Treatment and Assessment, Naming, Word-Class Phenomena, Agrammatism, Dementia, and Methodology* of the *Academy of Aphasia 34th Annual Meeting—London, England*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Belletti, A., & Rizzi, L. (1988). Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, *6*(3), 291–352.

Beretta, A., & Campbell, C. (2001). Psychological verbs and the double-dependency hypothesis. *Brain and Cognition*, *46*(1–2), 42–46.

Bickel, B., Witzlack-Makarevich, A., Choudhary, K. K., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2015). The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLoS ONE*, *10*(8), 1–22.

Bidgood, A., Pine, J. M., Rowland, C. F., & Ambridge, B. (2020). Syntactic representations are both abstract and semantically constrained: Evidence from children's and adults' comprehension and production/priming of the English passive. *Cognitive Science*, *44*(9), e12892.

Black, M., Nickels, L., & Byng, S. (1991). Patterns of sentence processing deficit: Processing simple sentences can be a complex matter. *Journal of Neurolinguistics*, *6*(2), 79–101.

Bornkessel, I., McElree, B., Schlesewsky, M., & Friederici, A. D. (2004). Multi-dimensional contributions to garden path strength: Dissociating phrase structure from case marking. *Journal of Memory and Language*, *51*(4), 495–522.

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, *113*(4), 787–821.

Bornkessel, I., Schlesewsky, M., & Friederici, A. D. (2002). Beyond syntax: Language-related positivities reflect the revision of hierarchies. *NeuroReport*, *13*(3), 361–364.

——— (2003). Eliciting thematic reanalysis effects: The role of syntax-independent information during parsing. *Language and Cognitive Processes*, *18*(3), 269–298.

Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y., & Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies during language comprehension. *NeuroImage*, *26*(1), 221–233.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. *Language and Linguistics Compass*, *3*(1), 19–58.

Brennan, J., & Pylkkänen, L. (2010). Processing psych verbs: Behavioural and MEG measures of two different types of semantic complexity. *Language and Cognitive Processes*, *25*(6), 777–807.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavioral Research Methods*, *41*, 977–990.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

——— (2018). Advanced Bayesian multilevel modeling with R package brms. *The R Journal*, *10*(1), 395–411.

——— (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54.

Carrithers, C. (1989). Syntactic complexity does not necessarily make sentences harder to understand. *Journal of Psycholinguistic Research*, *18*(1), 75–88.

Christianson, K. (2001). Optimality Theory in language production: The choice between direct and inverse in Odawa. *Linguistica Atlantica*, *23*, 93–125.

Christianson, K., & Cho, H. Y. (2009). Interpreting null pronouns (*pro*) in isolated sentences. *Lingua*, *119*(7), 989–1008.

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, *98*, 105–135.

Crain, S., Thornton, R., & Murasugi, K. (2009). Capturing the evasive passive. *Language Acquisition*, *16*(2), 123–133.

Cupples, L. (2002). The structural characteristics and on-line comprehension of experiencer-verb sentences. *Language and Cognitive Processes*, *17*(2), 125–162.

Darby, J. (2016). Assessing agentivity and eventivity in object-experiencer verbs: The role of processing. Paper presented at the 38th Deutschen Gesellschaft für Sprachwissenschaft, 26 February 2016.

Do, M. L., & Kaiser, E. (2019). The syntax-to-semantics mapping in real-time language production: A view from psych verbs. *LSA 93*. Talk at the 93rd Annual Meeting of the Linguistic Society of America, 6 January 2019, Sheraton New York Times Square.

——— (2022). Sentence formulation is easier when thematic and syntactic prominence align: Evidence from psych verbs. *Language, Cognition and Neuroscience*, *37*(5), 648–670.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619.

Dröge, A., Maffongelli, L., & Bornkessel-Schlesewsky, I. (2014). Luigi piace a Laura? Electrophysiological evidence for thematic reanalysis with Italian dative object experiencer verbs. In A. Bachrach, I. Roy, & L. Stockall (Eds.) *Structuring the Argument: Multidisciplinary Research on Verb Argument Structure*, vol. 10 of *Language Faculty and Beyond: Internal and External Variation in Linguistics*, (pp. 83–118). Amsterdam: John Benjamins.

Drummond, A. (2011). *IbexFarm (Version 0.3.9) [Software]*. Now defunct. Previously avaiable at `https://www.spellout.net/ibexfarm/`.

Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, *33*, 715–736.

——— (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83.

Fillmore, C. J. (1968). The case for case. In E. Bach, & R. T. Harms (Eds.) *Universals in Linguistic Theory*. New York: Holt, Rinehart & Winston.

Fox, D., & Grodzinsky, Y. (1998). Children's passive: A view from the *by*-phrase. *Linguistic Inquiry*, *29*(2), 311–332.

Gale, W. A., & Church, K. W. (1990). Estimation procedures for language context: Poor estimates are worse than none. In K. Momirović, & V. Mildner (Eds.) *Compstat: Proceedings in Computational Statistics, 9th Symposium Held at Dubrovnik, Yugoslavia, 1990*, (pp. 69–74). Heidelberg: Physica-Verlag Heidelberg.

Gattei, C. A., Alvarez, F., París, L. A., Wainselboim, A. J., Sevilla, Y., & Shalom, D. (2022). Why bother? What our eyes tell about linking during the comprehension of psych verb (non) causative constructions. *Glossa Psycholinguistics*, *1*(1), 1–38.

Gattei, C. A., Dickey, M. W., Wainselboim, A. J., & París, L. (2015a). The thematic hierarchy in sentence comprehension: A study on the interaction between verb class and word order in Spanish. *The Quarterly Journal of Experimental Psychology*, *68*(10), 1981–2007.

Gattei, C. A., Sevilla, Y., Tabullo, Á. J., Wainselboim, A. J., París, L. A., & Shalom, D. E. (2017). Prominence in Spanish sentence comprehension: An eye-tracking study. *Language, Cognition and Neuroscience*, *33*(5), 1–21.

Gattei, C. A., Tabullo, A., París, L., & Wainselboim, A. J. (2015b). The role of prominence in Spanish sentence comprehension: An ERP study. *Brain & Language*, *150*, 22–35.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, *58*(2), 161–187.

——— (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*(1), 1–23.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(20), 8051–8056.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., & Bergen, L. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Givón, T. (1984). *Syntax: A Functional-Typological Introduction, Volume I*. Philadelphia: John Benjamins.

Gómez-Vidal, B., Arantzeta, M., Laka, J. P., & Laka, I. (2022). Subjects are not all alike: Eye-tracking the agent preference in Spanish. *PLoS ONE*, *17*(8), 1–24.

Gordon, P., & Chafetz, J. (1990). Verb-based versus class-based accounts of actionality effects in children's comprehension of passives. *Cognition*, *36*, 227–254.

Grimshaw, J. (1990). *Argument Structure*. Cambridge, MA: MIT Press.

Grodzinsky, Y. (1995). Trace deletion, θ-roles and cognitive strategies. *Brain and Language*, *51*(3), 469–497.

Grodzinsky, Y., Pierce, A., & Marakovitz, S. (1991). Neuropsychological reasons for a transformational analysis of verbal passive. *Natural Language & Linguistic Theory*, *9*(3), 431–453.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, *142*(3), 880–905.

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, *175*, 36–52.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, (pp. 1–8). USA: Association for Computational Linguistics. Available at `https://doi.org/10.3115/1073336.1073357`

Hammerly, C. (2020). *Person-based Prominence in Ojibwe*. Ph.D. thesis, University of Massachusetts Amherst.

Hammerly, C., Staub, A., & Dillon, B. (2022). Person-based prominence guides incremental interpretation: Evidence from obviation in Ojibwe. *Cognition*, *225*, 105122.

Hartshorne, J. K., Pogue, A., & Snedeker, J. (2015). Love is hard to understand: The relationship between transitivity and caused events in the acquisition of emotion verbs. *Journal of Child Language*, *42*(3), 467–504.

Hirsch, C., & Wexler, K. (2006). Children's passives and their *resulting* interpretation. In K. U. Deen, J. Nomura, B. Schulz, & B. D. Schwarz (Eds.) *The Proceedings of the Inaugural Conference on Generative Approaches to Language Acquisition—North America*, vol. 1 of *University of Connecticut Occassional Papers in Linguistics #4*, (pp. 125–136). Cambridge, MA: MIT Working Papers in Linguistics.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2023). Surprisal does not explain syntactic disambiguation difficulty: Evidence from a large-scale benchmark. Available at psyarxiv.com/z38u6.

Huber, E., Sauppe, S., Isasi-Isasmendi, A., Bornkessel-Schlesewsky, I., Merlo, P., & Bickel, B. (2023). Surprisal from language models can predict ERPs in processing predicate-argument structure only if enriched by an Agent Preference principle. *Neurobiology of Language*, (pp. 1–68). Available at doi.org/10.1162/nol_a_00121.

Isasi-Isasmendi, A., Sauppe, S., Andrews, C., Laka, I., Meyer, M., & Bickel, B. (2023). Incremental sentence processing is guided by a preference for agents: EEG evidence from Basque. *Language, Cognition and Neuroscience*, (pp. 1–22). Available at https://doi.org/10.1080/23273798.2023.2250023.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*, vol. 2 of *Studies in Linguistics*. Cambridge, MA: MIT Press.

——— (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jeffreys, H. (1939). *Theory of Probability*. Oxford: Clarendon Press.

Kretzschmar, F., Bornkessel-Schlesewsky, I., Staub, A., Roehm, D., & Schlesewsky, M. (2012). Prominence facilitates ambiguity resolution: On the interaction between referentiality, thematic roles and word order in syntactic reanalysis. In M. J. A. Lamers, & P. de Swart (Eds.) *Case, Word Order and Prominence: Interacting Cues in Language Production and Comprehension*, vol. 40 of *Studies in Theoretical Psycholinguistics*, (pp. 239–271). Dordrecht: Springer.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Lamers, M. J. A. (2007). Verb type, animacy and definiteness in grammatical function disambiguation. In B. Los, & M. van Koppen (Eds.) *Linguistics in the Netherlands 2007*, vol. 24 of *Algemene Vereniging voor Tallwetenschap Publications*, (pp. 125–137). Amsterdam: John Benjamins.

Landau, I. (2010). *The Locative Syntax of Experiencers*, vol. 51 of *Linguistic Inquiry Monographs*. Cambridge: MIT Press.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, Illinois: University of Chicago Press.

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

——— (2008b). A noisy channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Thirteenth Conference on Empirical Methods in Natural Language Processing*, (pp. 234–243). Stroudsburg, PA: Association for Computational Linguistics.

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, *44*, 27–46.

Liter, A., Huelskamp, T., Weerakoon, S., & Munn, A. (2015). What drives the Maratsos effect, agentivity or eventivity? Poster presented at the 40th Annual Boston University Conference on Language Development.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*(226), 1–16.

Manouilidou, C., & de Almeida, R. G. (2013). Processing correlates of verb typologies: Investigating internal structure and argument realization. *Linguistics*, *51*(4), 767–792.

Maratsos, M., & Abramovitch, R. (1975). How children understand full, truncated, and anomalous passives. *Journal of Verbal Learning and Verbal Behavior*, *14*, 145–157.

Maratsos, M., Fox, D. E. C., Becker, J. A., & Chalkley, M. A. (1985). Semantic restrictions on children's passives. *Cognition*, *19*, 167–191.

Mateu, V. (2022). Object *wh*-questions with psych verbs are easy in child Spanish. In Y. Gong, & F. Kpogo (Eds.) *Proceedings of the 46th Annual Boston University Conference on Language Development*, (pp. 524–537). Somerville, MA: Cascadilla Press.

Meng, M., & Bader, M. (2021). Does comprehension (sometimes) go wrong for noncanonical sentences? *Quarterly Journal of Experimental Psychology*, *74*(1), 1–28.

Messenger, K., Branigan, H. P., McLean, J. F., & Sorace, A. (2012). Is young children's passive syntax semantically constrained? evidence from syntactic priming. *Journal of Memory and Language*, *66*, 568–587.

Minkoff, S. (1994). *How some so-called "thematic roles" that select animate arguments are generated, and how these roles inform binding and control*. Ph.D. thesis, MIT.

Momma, S., & Phillips, C. (2018). The relationship between parsing and generation. *Annual Review of Linguistics*, *4*(1), 233–254.

Nguyen, E. (2015). Felicity of the *by*-phrase is not enough: Re-evaluating O'Brien et al. (2006). University of Connecticut, Storrs, Unpublished m.s.

——— (2021). *The Predictive Power of Lexical Semantics on the Acquisition of Passive Voice*. Ph.D. thesis, University of Connecticut.

Nguyen, E., & Pearl, L. S. (2018). Do you really mean it? Linking lexical semantic profiles and age of acquisition for the English passive. In W. G. Bennett, L. Hracs, & D. R. Storoshenko (Eds.) *Proceedings of the 35th West Coast Conference on Formal Linguistics*, (pp. 288–295). Somerville, MA: Cascadilla Press.

——— (2019). Using developmental modeling to specify learning and representation of the passive in English children. In M. M. Brown, & B. Dailey (Eds.) *Proceedings of the 43rd Annual Boston University Conference on Language Development*, vol. 2, (pp. 469–482). Somerville, MA: Cascadilla Press.

Nguyen, E., & Pearl, S., Lisa (2021). The link between lexical semantic features and children's comprehension of English verbal *be*-passives. *Language Acquisition*, *28*(4), 433–450.

Nicenboim, B., Schad, D., & Vasishth, S. (2022). An introduction to Bayesian data analysis for cognitive science. Available online at vasishth.github.io/bayescogsci/book.

O'Brien, K., Grolla, E., & Lillo-Martin, D. (2006). Long passives are understood by young children. In D. Bamman, T. Magnitskaia, & C. Zaller (Eds.) *Proceedings of the 30th Annual Boston University Conference on Language Development*, vol. 2, (pp. 441–451). Somerville, MA: Cascadilla Press.

Orfitelli, R. M. (2012). *Argument Intervention in the Acquisition of A-movement*. Ph.D. thesis, University of California, Los Angeles.

Paolazzi, C. L. (2018). *Processing Passive Sentences in English: A Cross-methodological Study*. Ph.D. thesis, University College London.

Paolazzi, C. L., Grillo, N., Alexiadou, A., & Santi, A. (2019). Passives are not hard to interpret but hard to remember: Evidence from online and offline studies. *Language, Cognition and Neuroscience*, *34*(8), 991–1015.

Paolazzi, C. L., Grillo, N., Cera, C., Karageorgou, F., Bullman, E., Chow, W. Y., & Santi, A. (2021). Eyetracking while reading passives: An event structure account of difficulty. *Language, Cognition and Neuroscience*, *37*(2), 135–153.

Pearl, L. S., & Sprouse, J. (2019). Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. *Language*, *95*(4), 583–611.

——— (2021). The acquisition of linking theories: A Tolerance and Sufficiency Principle approach to deriving UTAH and rUTAH. *Language Acquisition*, *28*(3), 294–325.

Perlmutter, D., & Postal, P. (1984). The 1-advancement exclusiveness law. In D. Perlmutter, & C. Rosen (Eds.) *Studies in Relational Grammar 2*, (pp. 81–125). Chicago: University of Chicago Press.

Perovic, A., Vuksanović, J., Petrović, B., & Avramović-Ilić, I. (2014). The acquisition of passives in Serbian. *Applied Psycholinguistics*, *35*(1), 1–26.

Pesetsky, D. (1995). *Zero Syntax: Experiencers and Cascades*. Current Studies in Linguistics. Cambridge, MA: MIT Press.

Phillips, C. (1996). *Order and Structure*. Ph.D. thesis, MIT.

Piñango, M. M. (2000). Canonicity in Broca's sentence comprehension: The case of psychological verbs. In Y. Grodzinsky, L. Shapiro, & D. Swinney (Eds.) *Language and the Brain: Representation and Processing*, Foundations of Neuropsychology, (pp. 327–350). San Diego: Academic Press.

Pinker, S., Lebeaux, D. S., & Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. *Cognition*, *26*(3), 195–267.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at R-project.org.

Rappaport, M., & Levin, B. (1988). What to do with θ-roles. In W. Wilkins (Ed.) *Thematic Relations*, vol. 21 of *Syntax and Semantics*, (pp. 7–36). San Diego, CA: Academic Press.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. P. (2022). Large-scale evidence for logarithmic effects of word predictability on reading time. Available at psyarxiv.com/4hyna.

Slobin, D. I. (1966). Grammatical transformations and sentence comprehension in childhood and adulthood. *Journal of Verbal Learning and Verbal Behavior*, *5*(3), 219–227.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time in logarithmic. *Cognition*, *128*(3), 302–319.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86.

Temme, A., & Verhoeven, E. (2016). Verb class, case, and order: A crosslinguistic experiment on non-nominative experiencers. *Linguistics*, *54*, 769–813.

Thompson, C. K., & Lee, M. (2009). Psych verb production and comprehension in agrammatic Broca's aphasia. *Journal of Neurolinguistics*, *22*(4), 354–369.

Tiao, G. C., & Box, G. E. P. (1973). Some comments on "Bayes" estimators. *The American Statistician*, *27*(1), 12–14.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90.

Van Valin, R. D., Jr. (1990). Semantic parameters of split intransitivity. *Language*, *66*(2), 221–260.

Vani, P., Wilcox, E. G., & Levy, R. (2021). Using the interpolated maze task to assess incremental processing in English relative clauses. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, (pp. 1528–1534). Online: Cogntive Science Society.

Verhoeven, E. (2010). Agentivity and stativity in experiencer verbs: Implications for a typology of verb classes. *Linguistic Typology*, *14*(2–3), 213–251.

——— (2014). Thematic prominence and animacy asymmetries. evidence from a cross-linguistic production study. *Lingua*, *143*, 129–161.

——— (2015). Thematic asymmetries do matter! A corpus study of German word order. *Journal of German Linguistics*, *27*(1), 45–104.

de Villiers, J. G., & de Villiers, P. A. (1973). Development of the use of word order in comprehension. *Journal of Psycholinguistic Research*, *2*, 331–341.

Wagers, M., & Pendleton, E. (2016). Structuring expectation: Licensing animacy in relative clause comprehension. In K.-m. Kim, P. Umbal, T. Block, Q. Chan, T. Cheng, K. Finney, M. Katz, S. Nickel-Thompson, & L. Shorten (Eds.) *Proceedings of the 33rd West Coast Conference on Formal Linguistics*, (pp. 29–46). Somerville, MA: Cascadilla Press.

Wagers, M. W., Borja, M. F., & Chung, S. (2018). Grammatical licensing and relative clause parsing in a flexible word-order language. *Cognition*, *178*, 207–221.

Wilcox, E., & Vani, R., Pranali andH Levy (2021). A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (pp. 939–952). Online: Association for Computational Linguistics. Available at aclanthology.org/2021.acl-long.76.

Wilson, V. A. D., Zuberbühler, K., & Bickel, B. (2022). The evolutionary origins of syntax: Event cognition in nonhuman primates. *Science Advances*, *8*(25), 1–12.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. Available at doi.org/10.17605/OSF.IO/MD832.

A. Stimuli

We used the same items for experiments 1 and 2, with the exception that experiment 2 did not include any questions.

We do not list question conditions here. The questions for each item can be constructed by taking the embedded clause verb (e.g., *bother*), and placing it in the active and the passive voice to form two different question conditions per sentence condition, as follows: active question, *Who bothered someone?*; passive question, *Who was bothered?*. The possible answers to each question consisted of the two arguments of the embedded clause verb in both possible orders (e.g., *the actor the director* or *the director the actor*). Thus for each sentence condition (4 total), there were 4 question conditions crossing Question voice and Answer order independently of Sentence voice, leading to 16 conditions per item total in experiment 1.

Each item is numbered according to the list it occurred in. The orders were constructed such that participants would see each member of a pair of items in opposite conditions. For instance, if a participant saw item 1 in the active, subject experiencer condition, they would see item 2 in the passive, object experiencer condition. While we don't show the questions here, this was true of questions too, such that Question voice and Answer order were also always shown in the opposite conditions for paired items. Each odd-numbered item was in such a pair with the following even-numbered item. This ensured that participants saw equal numbers of items in each of the 16 total conditions in experiment 1, and each of the 4 total conditions in experiment 2.

We use A and P to abbreviate active and passive, and SE and OE to abbreviate subject experiencer and object experiencer, respectively.

1.	(a)	That's the actor that the director particularly appreciates because of his dramatic emotional bearing.	(A, SE)
	(b)	That's the actor that the director particularly bothers because of his dramatic emotional bearing.	(A, OE)
	(c)	That's the director that the actor is particularly appreciated by because of his dramatic emotional bearing.	(P, SE)
	(d)	That's the director that the actor is particularly bothered by because of his dramatic emotional bearing.	(P, OE)
2.	(a)	That's the park ranger that the tourist particularly appreciates because of his gung-ho attitude.	(A, SE)
	(b)	That's the tourist that the park ranger is particularly bothered by because of his gung-ho attitude.	(P, OE)
	(c)	That's the tourist that the park ranger is particularly appreciated by because of his gung-ho attitude.	(P, SE)
	(d)	That's the park ranger that the tourist particularly bothers because of his gung-ho attitude.	(A, OE)
3.	(a)	That's the student that the professor somewhat admires in spite of his know-it-all attitude.	(A, SE)
	(b)	That's the student that the professor somewhat frightens in spite of his know-it-all attitude.	(A, OE)
	(c)	That's the professor that the student is somewhat admired by in spite of his know-it-all attitude.	(P, SE)
	(d)	That's the professor that the student is somewhat frightened by in spite of his know-it-all attitude.	(P, OE)
4.	(a)	That's the player that the coach somewhat admires in spite of his over-eagerness.	(A, SE)
	(b)	That's the coach that the player is somewhat frightened by in spite of his over-eagerness.	(P, OE)
	(c)	That's the coach that the player is somewhat admired by in spite of his over-eagerness.	(P, SE)
	(d)	That's the player that the coach somewhat frightens in spite of his over-eagerness.	(A, OE)

5. (a) That's the musician that the conductor greatly adores due to his vigorous practice regimen. (A, SE)
   (b) That's the musician that the conductor greatly disappoints due to his vigorous practice regimen. (A, OE)
   (c) That's the conductor that the musician is greatly adored by due to his vigorous practice regimen. (P, SE)
   (d) That's the conductor that the musician is greatly disappointed by due to his vigorous practice regimen. (P, OE)

6. (a) That's the carpenter that the contractor greatly adores due to his workmanship. (A, SE)
   (b) That's the contractor that the carpenter is greatly disappointed by due to his workmanship. (P, OE)
   (c) That's the contractor that the carpenter is greatly adored by due to his workmanship. (P, SE)
   (d) That's the carpenter that the contractor greatly disappoints due to his workmanship. (A, OE)

7. (a) That's the greengrocer that the tailor really fancies on account of their long conversations. (A, SE)
   (b) That's the greengrocer that the tailor really bores on account of their long conversations (A, OE)
   (c) That's the tailor that the greengrocer is really fancied by on account of their long conversations. (P, SE)
   (d) That's the tailor that the greengrocer is really bored by on account of their long conversations. (P, OE)

8. (a) That's the bellboy that the guest really fancies on account of his placid demeanor. (A, SE)
   (b) That's the guest that the bellboy is really bored by on account of his placid demeanor. (P, OE)
   (c) That's the guest that the bellboy is really fancied by on account of his placid demeanor. (P, SE)
   (d) That's the bellboy that the guest really bores on account of his placid demeanor. (A, OE)

9. (a) That's the student that the teacher somewhat tolerates despite his uncaring appearance. (A, SE)
   (b) That's the student that the teacher somewhat terrifies despite his uncaring appearance. (A, OE)
   (c) That's the teacher that the student is somewhat tolerated by despite his uncaring appearance. (P, SE)
   (d) That's the teacher that the student is somewhat terrified by despite his uncaring appearance. (P, OE)

10. (a) That's the concierge that the executive somewhat tolerates despite his meek disposition. (A, SE)
    (b) That's the executive that the concierge is somewhat terrified by despite his meek disposition. (P, OE)
    (c) That's the executive that the concierge is somewhat tolerated by despite his meek disposition. (P, SE)
    (d) That's the concierge that the executive somewhat terrifies despite his meek disposition. (A, OE)

11. (a) That's the stockbroker that the celebrity really envies because of his vast wealth. (A, SE)
    (b) That's the stockbroker that the celebrity really interests because of his vast wealth. (A, OE)
    (c) That's the celebrity that the stockbroker is really envied by because of his vast wealth. (P, SE)
    (d) That's the celebrity that the stockbroker is really interested by because of his vast wealth. (P, OE)

12. (a) That's the farmer that the shopkeeper really envies because of his land holdings. (A, SE)
    (b) That's the shopkeeper that the farmer is really interested by because of his land holdings. (P, OE)
    (c) That's the shopkeeper that the farmer is really envied by because of his land holdings. (P, SE)
    (d) That's the farmer that the shopkeeper really interests because of his land holdings. (A, OE)

13. (a) That's the daughter that the father particularly cherishes in spite of her lackadaisical attitude. (A, SE)
    (b) That's the daughter that the father particularly annoys in spite of her lackadaisical attitude. (A, OE)
    (c) That's the father that the daughter is particularly cherished by in spite of her lackadaisical attitude. (P, SE)
    (d) That's the father that the daughter is particularly annoyed by in spite of her lackadaisical attitude. (P, OE)

14. (a) That's the colleague that the scientist particularly cherishes in spite of his nonchalant demeanor. (A, SE)
    (b) That's the scientist that the colleague is particularly annoyed by in spite of his nonchalant demeanor. (P, OE)
    (c) That's the scientist that the colleague is particularly cherished by in spite of his nonchalant demeanor. (P, SE)
    (d) That's the colleague that the scientist particularly annoys in spite of his nonchalant demeanor. (A, OE)

15. (a) That's the barber that the stylist really dislikes due to his bold new approach. (A, SE)
    (b) That's the barber that the stylist really amazes due to his bold new approach. (A, OE)

(c) That's the stylist that the barber is really disliked by due to his bold new approach. (P, SE)

(d) That's the stylist that the barber is really amazed by due to his bold new approach. (P, OE)

16. (a) That's the painter that the artist really dislikes due to her freeflowing compositions. (A, SE)

(b) That's the artist that the painter is really amazed by due to her freeflowing compositions. (P, OE)

(c) That's the artist that the painter is really disliked by due to her freeflowing compositions. (P, SE)

(d) That's the painter that the artist really amazes due to her freeflowing compositions. (A, OE)

17. (a) That's the thief that the cop strongly loathes on account of their constant chases. (A, SE)

(b) That's the thief that the cop strongly irritates on account of their constant chases. (A, OE)

(c) That's the cop that the thief is strongly loathed by on account of their constant chases. (P, SE)

(d) That's the cop that the thief is strongly irritated by on account of their constant chases. (P, OE)

18. (a) That's the minister that the king strongly loathes on account of his poor judgement. (A, SE)

(b) That's the king that the minister is strongly irritated by on account of his poor judgement. (P, OE)

(c) That's the king that the minister is strongly loathed by on account of his poor judgement. (P, SE)

(d) That's the minister that the king strongly irritates on account of his poor judgement. (A, OE)

19. (a) That's the boss that the employee somewhat dreads despite his apparent cheerfulness. (A, SE)

(b) That's the boss that the employee somewhat depresses despite his apparent cheerfulness. (A, OE)

(c) That's the employee that the boss is somewhat dreaded by despite his apparent cheerfulness. (P, SE)

(d) That's the employee that the boss is somewhat depressed by despite his apparent cheerfulness. (P, OE)

20. (a) That's the editor that the journalist somewhat dreads despite his upbeat personality. (A, SE)

(b) That's the journalist that the editor is somewhat depressed by despite his upbeat personality. (P, OE)

(c) That's the journalist that the editor is somewhat dreaded by despite his upbeat personality. (P, SE)

(d) That's the editor that the journalist somewhat depresses despite his upbeat personality. (A, OE)

21. (a) That's the butcher that the monk greatly detests because of his unexpectedly bawdy behavior. (A, SE)

(b) That's the butcher that the monk greatly intrigues because of his unexpectedly bawdy behavior. (A, OE)

(c) That's the monk that the butcher is greatly detested by because of his unexpectedly bawdy behavior. (P, SE)

(d) That's the monk that the butcher is greatly intrigued by because of his unexpectedly bawdy behavior. (P, OE)

22. (a) That's the auditor that the businessman greatly detests because of his enormous wealth. (A, SE)

(b) That's the businessman that the auditor is greatly intrigued by because of his enormous wealth. (P, OE)

(c) That's the businessman that the auditor is greatly detested by because of his enormous wealth. (P, SE)

(d) That's the auditor that the businessman greatly intrigues because of his enormous wealth. (A, OE)

23. (a) That's the author that the reviewer strongly reveres in spite of his frequent trips to Europe. (A, SE)

(b) That's the author that the reviewer strongly repulses in spite of his frequent trips to Europe. (A, OE)

(c) That's the reviewer that the author is strongly revered by in spite of his frequent trips to Europe. (P, SE)

(d) That's the reviewer that the author is strongly repulsed by in spite of his frequent trips to Europe. (P, OE)

24. (a) That's the carpenter that the stonemason strongly reveres in spite of his mediocrity. (A, SE)

(b) That's the stonemason that the carpenter is strongly repulsed by in spite of his mediocrity. (P, OE)

(c) That's the stonemason that the carpenter is strongly revered by in spite of his mediocrity. (P, SE)

(d) That's the carpenter that the stonemason strongly repulses in spite of his mediocrity. (A, OE)

25. (a) That's the priest that the reporter particularly idolizes due to his behavior at press conferences. (A, SE)

(b) That's the priest that the reporter particularly unnerves due to his behavior at press conferences. (A, OE)

(c) That's the reporter that the priest is particularly idolized by due to his behavior at press conferences. (P, SE)

(d) That's the reporter that the priest is particularly unnerved by due to his behavior at press conferences. (P, OE)

26. (a) That's the butler that the millionaire particularly idolizes due to his distinctive mannerisms. (A, SE)
(b) That's the millionaire that the butler is particularly unnerved by due to his distinctive mannerisms. (P, OE)
(c) That's the millionaire that the butler is particularly idolized by due to his distinctive mannerisms. (P, SE)
(d) That's the butler that the millionaire particularly unnerves due to his distinctive mannerisms. (A, OE)

27. (a) That's the scientist that the artist strongly distrusts on account of his being late for meetings. (A, SE)
(b) That's the scientist that the artist strongly perturbs on account of his being late for meetings. (A, OE)
(c) That's the artist that the scientist is strongly distrusted by on account of his being late for meetings. (P, SE)
(d) That's the artist that the scientist is strongly perturbed by on account of his being late for meetings. (P, OE)

28. (a) That's the drummer that the guitarist strongly distrusts on account of his never-ending primadonnism. (A, SE)
(b) That's the guitarist that the drummer is strongly perturbed by on account of his never-ending primadonnism. (P, OE)
(c) That's the guitarist that the drummer is strongly distrusted by on account of his never-ending primadonnism. (P, SE)
(d) That's the drummer that the guitarist strongly perturbs on account of his never-ending primadonnism. (A, OE)

29. (a) That's the lawyer that the judge greatly esteems despite his conduct in the courtroom. (A, SE)
(b) That's the lawyer that the judge greatly appalls despite his conduct in the courtroom. (A, OE)
(c) That's the judge that the lawyer is greatly esteemed by despite his conduct in the courtroom. (P, SE)
(d) That's the judge that the lawyer is greatly appalled by despite his conduct in the courtroom. (P, OE)

30. (a) That's the florist that the gardener greatly esteems despite her being of average skill. (A, SE)
(b) That's the gardener that the florist is greatly appalled by despite her being of average skill. (P, OE)
(c) That's the gardener that the florist is greatly esteemed by despite her being of average skill. (P, SE)
(d) That's the florist that the gardener greatly appalls despite her being of average skill. (A, OE)

31. (a) That's the client that the consultant really respects because of his out of the box thinking. (A, SE)
(b) That's the client that the consultant really impresses because of his out of the box thinking. (A, OE)
(c) That's the consultant that the client is really respected by because of his out of the box thinking. (P, SE)
(d) That's the consultant that the client is really impressed by because of his out of the box thinking. (P, OE)

32. (a) That's the agent that the rock star really respects because of his unmatched abilities. (A, SE)
(b) That's the rock star that the agent is really impressed by because of his unmatched abilities. (P, OE)
(c) That's the rock star that the agent is really respected by because of his unmatched abilities. (P, SE)
(d) That's the agent that the rock star really impresses because of his unmatched abilities. (A, OE)

B. Corpus study

In order to determine the potential effect of frequency on our results, we conducted a corpus study, following Gennari & MacDonald (2009). According to Surprisal Theory (Gennari & MacDonald 2009; Gibson et al. 2013, 2019; Levy 2008a,b; MacDonald 2013; Shain et al. 2022; Smith & Levy 2013; Vani et al. 2021; Wilcox & Vani 2021; see also Huang et al. 2023 and Huber et al. 2023), the human sentence parser is chiefly influenced by its linguistic experience. The more commonly encountered a particular configuration is, the more experience a comprehender has understanding it, and the easier it is to comprehend (Gennari & MacDonald 2009; MacDonald 2013).

If processing is primarily influenced by frequency, as in Surprisal Theory, another explanation of our results might be possible depending on how frequent particular configurations are. For instance, suppose that object experiencer verbs are more frequently produced in the passive voice than subject experiencer verbs are (e.g. Ferreira 1994). Under Surprisal Theory, comprehenders' greater experience with passive object experiencer verbs could facilitate their processing in such configurations relative to passive subject experiencer verbs. The opposite might hold for active sentences.

To address this possibility, we coded frequency as the relative frequency of our subject- and object-experiencer verbs in the active and passive voices. We estimated these values using the SUBTLEX corpus of English-language subtitles, as this corpus has been shown to better predict reading times than other corpora (Brysbaert & New 2009). We extracted the first 150 lines containing any form of each verb from a text file that had lines from SUBTLEX in a pre-randomized order. Some of our verbs did not occur at least 150 times; for these verbs, we used all occurrences. Verbs that did not occur in at least 150 lines were *revere* (99, subjexp), *detest* (87, subjexp), *repulse* (55, objexp), *unnerve* (43, objexp), *distrust* (39, subjexp), *idolize* (28, subjexp), and *perturb* (22, objexp); thus four subject experiencer verbs and three object experiencer verbs had fewer data points with which we were able to gather frequency information than our other verbs.

Lines were manually coded according to whether the verb occurred in active or passive voice by the first author. In many cases, passives of object experiencer verbs were ambiguous between verbal and adjectival passive readings; in these cases, the verb was coded as occurring in passive voice, with a flag that the passive may have been an adjectival rather than a verbal passive. We note that according to alignment hypotheses that make reference to thematic roles, it should not matter whether a passive of an object experiencer verb is verbal or adjectival in nature, because in both cases it displays the same linking pattern (i.e., the subject of the predicate is the experiencer, and any oblique is the theme). Nonetheless, we coded this as well to ensure that any results of our analyses did not crucially depend on how these potentially ambiguous forms were categorized. We reasoned that using both ways of quantifying frequency would allow us to use whichever measure would be the strongest possible predictor of the behavior we observed. Nominal uses (e.g., *respect* as in *You have his respect*) were excluded, as were clearly irrelevant homophonic uses (e.g., *appreciate* in its financial sense). These data (with our coding) are available at osf.io/jq62t.

Table 1 reports the total counts and percentages of active and passive voice found with each category of verb, to give a birds-eye view of our findings.

	including probable adjectives		excluding probable adjectives	
	Active	Passive	Active	Passive
Subject experiencer	1277 (95.01%)	67 (4.99%)	1277 (96.45%)	47 (3.55%)
Object experiencer	727 (51.74%)	678 (48.26%)	423 (83.27%)	85 (16.73%)

Table 1: Count and % of voice usage by Verb type in our corpus study

Generally, our subject experiencer verbs overwhelmingly occur in active voice, no matter how ambiguous tokens are treated. When potential adjectival passives are included in our counts, object experiencer verbs occur almost equally in active and passive voice. When the count is restricted to unambiguous verbal passives, however, object experiencers also occur predominantly in the active voice. The results for object experiencer verbs including probable adjectives are similar to those reported in study 4 of Gennari & MacDonald (2009), who observed that object experiencer verbs with two animate arguments occurred in passives around 45% of the time (Gennari & MacDonald 2009).[8]

[8]However, note that unlike Gennari & MacDonald (2009), we did not restrict our frequency counts to a particular structural context, such as relative clauses. This is because, unlike in their study, we used the *by*-phrase argument as the

Next we considered the relative frequency of active and passive structures across each verb. Figures 4 and 5 show the percentages (on the *y*-axis) and counts (above each bar) including and excluding probable adjectives, respectively.

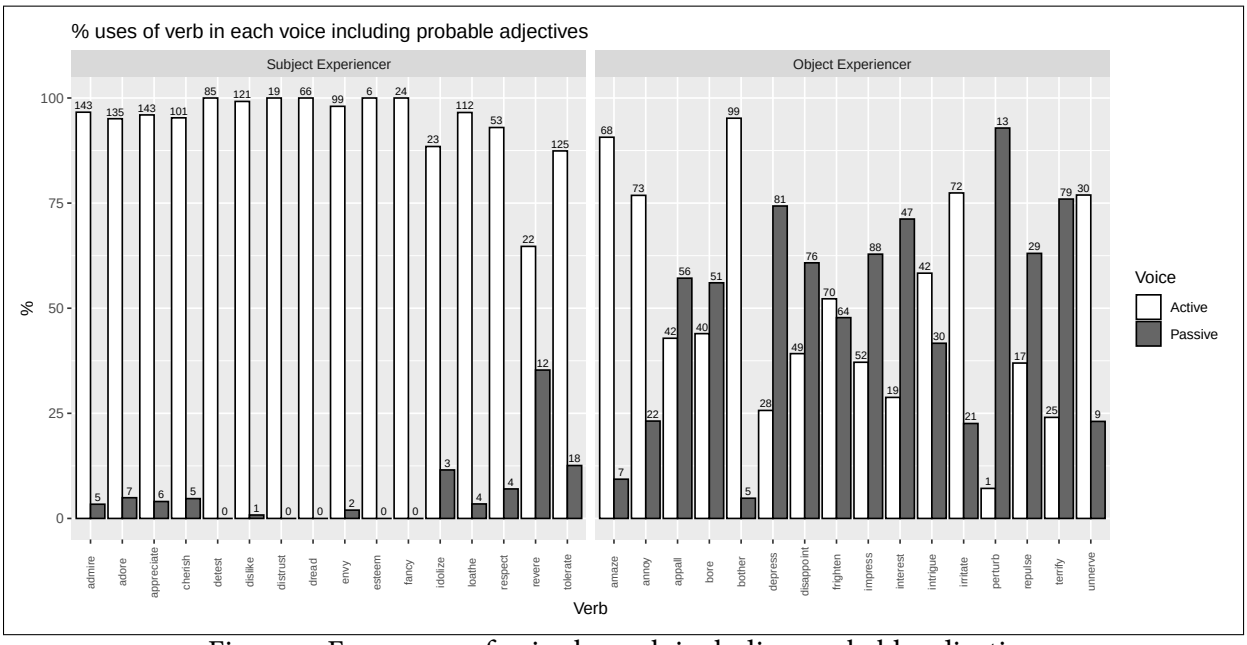


Figure 4: Frequency of voice by verb including probable adjectives

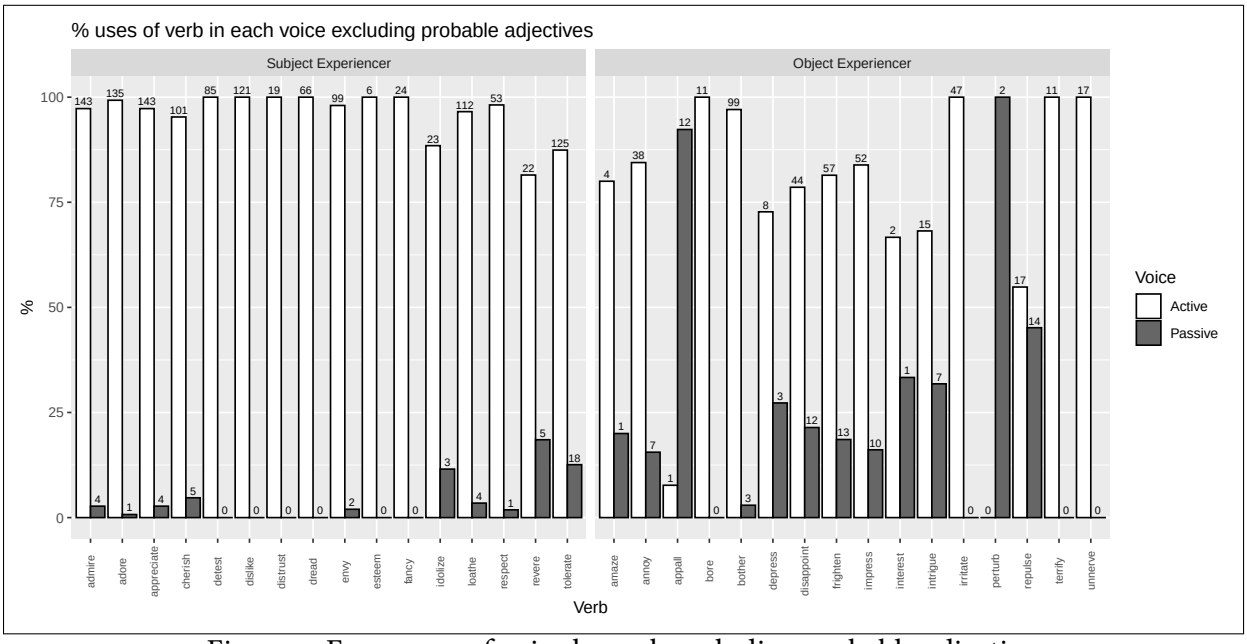


Figure 5: Frequency of voice by verb excluding probable adjectives

RC head in passives, as in *the director that the actor was really appreciated by*. Such structures are exceedingly infrequent for both subject and object experiencer verbs, and we found several of our verbs (perhaps all, though we did not check exhaustively) had no occurrences in this particular structure. However, we consider this move justified given discussion by Gennari & MacDonald (2009), who find that overall voice proportions predicted RC voice proportions well.

For subject experiencer verbs, we observe relatively little variation across verbs: all verbs in our sample occur much more often in active voice than in passive voice. In contrast, object experiencer verbs show a more heterogeneous profile; some occur much more frequently in the active voice, such as *annoy*, while others occur more frequently in the passive voice, such as *appall*. When probable adjectives are excluded from the counts, only two verbs—*appall* and *perturb*—show a clear passive bias.

Overall, our corpus study reveals that subject experiencer verbs occur much more frequently in active voice than in passive voice, no matter how potentially adjectival passives are counted. Object experiencer verbs, however, show a more mixed profile, either showing equal biases towards active and passive voice, or a bias towards active voice. We took the numerical results from this study and used them to examine the possible effects of frequency in Appendix C.

C. Frequency analysis and model comparisons

In order to account for a possible account of our effects in experiments 1 and 2 in terms of frequency (e.g., Surprisal Theory: Gennari & MacDonald 2009; Gibson et al. 2013, 2019; Hale 2001; Levy 2008a,b; MacDonald 2013; Shain et al. 2022; Smith & Levy 2013; Vani et al. 2021; Wilcox & Vani 2021 see also Huang et al. 2023 and Huber et al. 2023), we here present the results of models we fit including frequency as a predictor. We also present the results of model comparisons conducted to address the question of whether models using frequency as a predictor fared better than models not using frequency (as in the strongest possible version of Surprisal Theory; Levy 2008a; Smith & Levy 2013), or whether there were effects of frequency beyond the effects of the categorical predictors (Verb type and Sentence voice) we considered.

Gennari & MacDonald (2009) similarly used frequency (as estimated by counts obtained from a corpus study) to predict reading times for object experiencer verbs. However, while Gennari & MacDonald (2009) used proportions for their regression analyses, we instead used a log odds-based measure to account for large proportional changes in our predictors. We smoothed the raw counts for each verb and voice combination to account for zeros and the relatively low numbers of certain verbs occurring in certain voices by adding 0.5 to each count (the Expected Likelihood Estimate: Gale & Church 1990; Tiao & Box 1973). We then took the log odds ratio of the proportion of smoothed active/passive uses compared to the proportion of smoothed passive/active uses for each individual verb, as in (9),

$$\ln\left(\frac{n_{\text{Verb,Voice}_i} + 0.5}{n_{\text{Verb,Voice}_j} + 0.5}\right) \tag{9}$$

where $n_{\text{Verb,Voice}_i}$ is the number of observations of a particular verb in Voice$_i$, i is the voice the verb occurred in for that presentation of the item (Active or Passive), and j is the opposite voice compared to i. We refer to this value as the Sentence voice log odds. Thus, the Sentence voice log odds is the smoothed log odds of encountering the verb in the voice observed on a particular trial, compared to the unobserved voice, as based on the sentences we extracted from our corpus. Each verb is thus associated with two of these values (one for active uses and one for passive uses), which we used as a predictor for the corresponding conditions for each item. This meant that our frequency measure included information related to both the Sentence voice and Verb of each presentation of each item (and is thus more specific to particular items and conditions than the Verb type predictor). For the regression analyses, we used these values as a predictor of reading times and question accuracy.

C.1 Experiment 1

C.1.1 Voice frequency models

In addition to the models using only categorical predictors (Sentence voice, Verb type, and Question voice) described in the main text, we additionally fit linear models using only frequency information for the critical verb and spillover word, as well as a logistic model for question accuracy. The frequency measure we used was the log odds of the verb occurring in the voice (active or passive) that it appeared in in a particular presentation of a sentence as described above (we also fit models that included Question voice log odds as well as Sentence voice log odds for question accuracy). This was defined using the smoothed proportions including probable adjectives that we determined from the corpus study as described in section B and defined as in (9). We included probable adjectives in these and all subsequent models using frequency predictors, because those values had the best chance of the explaining the pattern of results we found.

As for the models presented in the main text, we fit both frequentist and Bayesian regressions using the same model specifications (with the exception that we had to simplify the random effects structures of the frequentist models to permit convergence). Results were comparable.

We can only obtain an imperfect estimate for our frequency-based predictor. That is, our estimate of frequency is the Sentence voice log odds of a verb-voice combination, as determined by our corpus study. But any corpus is necessarily incomplete, and thus our estimate is noisy. For this reason, the Bayesian models we fit take into account the error in our measurement (the standard error of the log odds ratio); see Nicenboim et al. (2022 ch. 13.2) for details. For our frequentist models (included with the supplementary materials), linear and logistic mixed effects measurement-error models are not currently supported in R to our knowledge, so we instead just used the log odds ratio directly as a predictor.

Plots showing the regression of means of our dependent variables on frequency are in figures 6 and 7. Note that while we took into account measurement error in the Sentence voice log odds (x-axis) in the models we fit, we do not plot error bars on the x-axis for reasons of visual clarity. Each point shows the mean value of a particular measure for one of the 32 verbs in a particular voice. For RT measures, a frequency-based approach predicts a line with a negative slope, since if a verb occurs in a particular voice more often, it should be faster to read in that voice. For questions, a frequency-based analysis predicts a line with a positive slope, since participants should respond accurately more often to a verb in a particular voice when they are more likely to encounter it in that voice.

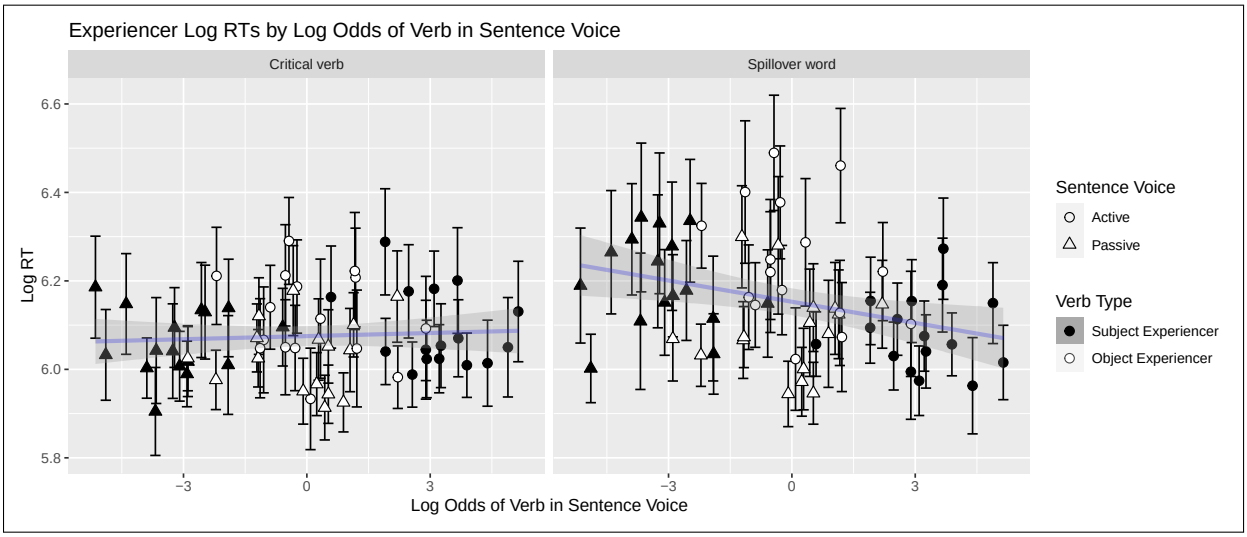


Figure 6: Regression of critical verb and spillover word mean RTs on frequency (Experiment 1)

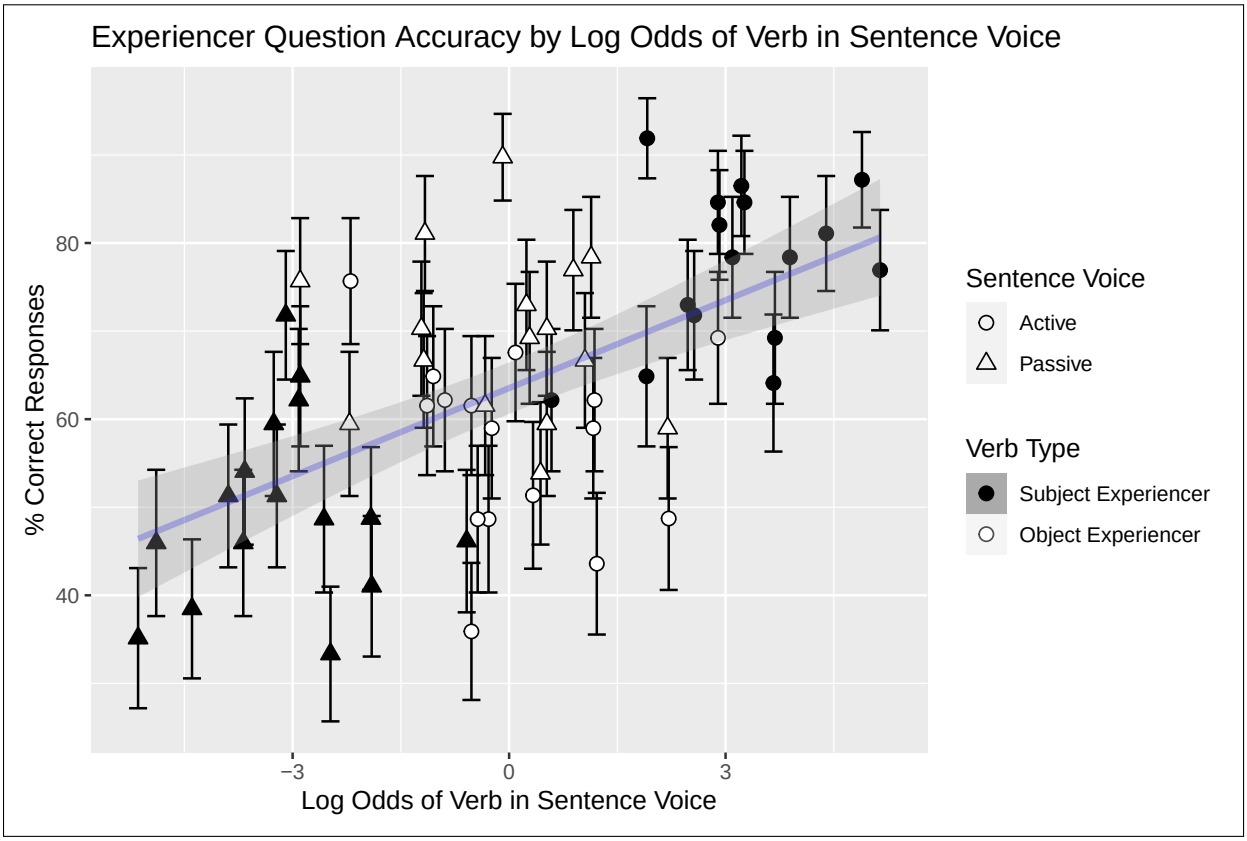


Figure 7: Regression of question accuracy on Sentence voice log odds (Experiment 1)

Critical verb There was no observed relationship between frequency and reading times at the critical verb (95% CI: $[-0.01, 0.02]$)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	6.07	0.05	5.98	6.17
Sentence voice log odds	0.00	0.01	−0.01	0.02

Table 1: Frequency model for critical verb (Experiment 1)

Spillover word At the spillover word, shorter RTs were associated with higher frequency (95% CI: $[-0.03, -0.01]$) In other words, participants read the spillover word faster when the verb occurred in its more frequent voice than when it occurred in its less frequent voice.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	6.15	0.05	6.05	6.26
Sentence voice log odds	−0.02	0.01	−0.03	−0.01

Table 2: Frequency model for spillover word (Experiment 1)

Question responses For question accuracy models using frequency as a predictor, there were two ways in which frequency might matter. This was because we manipulated Question voice independently of Sentence voice. For this reason, we considered a model that solely used Sentence voice frequency as a predictor of accuracy, as well as a model that predicted accuracy using the interaction of Sentence voice frequency and Question voice frequency. (We did not fit models that used only Question voice frequency as a predictor.) However, we found no relationship between Question voice frequency and accuracy in any model that included Question voice frequency as a predictor. For this reason, we do not present these models.

The question accuracy model using only Sentence voice frequency as a predictor found an effect of frequency (95% CI: $[0.08, 0.20]$) This reflected that when a verb occurred in its more frequent voice in the sentence, participants were more likely to answer the comprehension question correctly than when the verb occurred in its less frequent voice.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	0.60	0.09	0.42	0.78
Sentence voice log odds	0.15	0.03	0.08	0.20

Table 3: Sentence voice frequency model for question accuracy (Experiment 1)

C.1.2 Voice, Verb type, and Voice frequency models

In the examination of the models using only Sentence (and Question) voice and Verb type as predictors for experiment 1 (presented in the main text), we observed relationships between our categorical predictors and reading times and accuracy. However, in the frequency analyses in section C.1.1, we found a relationship between these predictors and frequency for the spillover word. In this section, we consider models regressing our measures on the interaction of our categorical predictors

as well as on frequency (and the interaction of different frequency measures for questions). This helps us determine whether both sets of predictors are relevant, and formed the basis for model comparisons in section C.1.4.

We did not consider models where frequency interacted with the categorical predictors, because frequency was already defined in terms of these predictors (i.e., the Sentence voice log odds is defined for a specific verb—which naturally has a single corresponding Verb type—and a particular voice; see (9)). Thus, it does not make sense to consider an interaction of the categorical predictors with frequency, as these are already baked into the definition of our frequency measure. As in the models using only frequency as a predictor, our Bayesian models that included both frequency and categorical predictors took into account measurement error in our estimate of frequency.

Critical verb The model fit using categorical predictors as well as frequency for the critical verb showed a relationship between Sentence voice (95% CI: $[-0.14, -0.00]$) and RTs, which was consistent with the model fit using only categorical predictors as well as the model fit using only frequency. This reflected that participants read the critical verb more quickly on average in passive sentences than in active sentences, but that there was no observed relationship between frequency and RTs for the critical verb.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	6.08	0.05	5.98	6.17
Verb type	−0.00	0.02	−0.04	0.04
Sentence voice	−0.07	0.03	−0.14	−0.00
Verb type × Sentence voice	−0.02	0.06	−0.15	0.10
Sentence voice log odds	−0.01	0.01	−0.03	0.02

Table 4: Crossed model for critical verb with frequency (Experiment 1)

Spillover word The model containing the categorical predictors and frequency for the spillover word found no relationship between any main effect and RTs. Nevertheless, we observed a relationship between the interaction of Verb type and Sentence voice and RTs (95% CI: $[-0.35, -0.02]$) This was consistent with the model fit using only categorical predictors; but contrasted with the model fit using only frequency.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	6.15	0.05	6.05	6.25
Verb type	0.02	0.03	−0.03	0.07
Sentence voice	−0.06	0.04	−0.15	0.02
Verb type × Sentence voice	−0.18	0.08	−0.35	−0.02
Sentence voice log odds	−0.02	0.01	−0.04	0.01

Table 5: Crossed model for spillover word with frequency (Experiment 1)

We note that the net effect of adding the frequency predictor to the model was mostly to greatly

spread out the estimate of the posterior for the interaction of Verb type and Sentence voice compared to the model that used only categorical predictors, and to somewhat spread out the estimate of frequency compared to the models fit using frequency alone. As table 2 showed, the 95% CI for the interaction effect in the model using only the categorical predictors was $[-0.39, -0.14]$ (range of 0.25), while this range for the interaction effect in the model that added frequency was $[-0.35, 0.02]$ (range of 0.33). Though we will not note this for every model going forward, we found this general pattern to be consistent, with the estimates of the main effects being little affected by the addition of frequency, and the estimates of the interaction being much less precise.

We resolved the interaction by both Verb type and Sentence voice. The model nesting Verb type in Sentence voice showed an effect of Verb type in active sentences (95% CI: $[0.02, 0.21]$) reflecting the fact that participants read the spillover word more quickly in active sentences when the sentence contained a subject experiencer verb than when it contained an object experiencer verb. However, no relationship was observed between Verb type and RTs in passive sentences (95% CI: $[-0.17, 0.03]$) This contrasted with the Verb type nested in Sentence voice model that did not include frequency, as that model showed a relationship between Verb type and RTs in passive sentences.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Verb type (active)	0.11	0.05	0.02	0.21
Verb type (passive)	−0.07	0.05	−0.17	0.03

Table 6: Verb type nested in Sentence voice model for spillover word with frequency (Experiment 1)

The model nesting Sentence voice within Verb type showed an effect of Sentence voice in object experiencer sentences (95% CI: $[-0.23, -0.08]$) This reflected that participants read the spillover word of sentences containing object experiencer verb faster for passives than for actives. While this finding was consistent with the model fit using only categorical predictors, the lack of an effect of Sentence voice in sentences with subject experiencer verbs contrasted with the results of that model.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Sentence voice (subjexp)	0.02	0.07	−0.12	0.17
Sentence voice (objexp)	−0.16	0.04	−0.23	−0.08

Table 7: Sentence voice nested in Verb type model for spillover word with frequency (Experiment 1)

We speculate that this is because subject experiencer verbs all occurred most often in the active voice (cf. figures 4–5), which made it difficult for the model to disentangle the effects of frequency from the effects of Sentence voice, thereby spreading out the posterior distribution. In contrast, object experiencer verbs had a different range of voice preferences, which meant that it was easier for the model to estimate frequency effects apart from effects of Sentence voice.

Question responses The model regressing question accuracy on the categorical predictors and their interactions as well as Sentence voice frequency showed a main effect of Sentence voice (95% CI: $[-0.66, -0.08]$) and an interaction between Verb type and Sentence voice (95% CI: $[1.33, 2.37]$).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	0.66	0.10	0.47	0.86
Verb type	−0.05	0.13	−0.30	0.21
Sentence voice	−0.37	0.15	−0.66	−0.08
Question voice	0.21	0.11	−0.01	0.43
Verb type × Sentence voice	1.85	0.26	1.33	2.37
Verb type × Question voice	−0.24	0.23	−0.70	0.22
Sentence voice × Question voice	0.01	0.23	−0.44	0.47
Verb type × Sentence voice × Question voice	0.35	0.38	−0.40	1.10
Sentence voice log odds	0.01	0.04	−0.06	0.08

Table 8: Crossed model for question accuracy with Sentence voice frequency (Experiment 1)

This pattern of effects was consistent with the model regressing question accuracy on the categorical predictors alone, but contrasted with the model fit using Sentence voice frequency alone—as that model showed an effect of Sentence voice frequency.

We resolved the interaction by both Verb type and Sentence voice. All effects matched those found in the nested models fit using only the categorical predictors.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Verb type (active)	−1.01	0.19	−1.39	−0.63
Verb type (passive)	0.92	0.18	0.56	1.28
Verb type (active) × Question voice	−0.43	0.30	−1.01	0.16
Verb type (passive) × Question voice	−0.05	0.30	−0.62	0.54

Table 9: Verb type nested in Sentence Voice model for question accuracy with Sentence voice frequency (Experiment 1)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Sentence voice (subjexp)	−1.32	0.24	−1.79	−0.86
Sentence voice (objexp)	0.59	0.15	0.30	0.89
Sentence voice (subjexp) × Question voice	−0.18	0.28	−0.72	0.37
Sentence voice (objexp) × Question voice	0.22	0.34	−0.44	0.89

Table 10: Sentence Voice nested in Verb type model for question accuracy with Sentence voice frequency (Experiment 1)

C.1.3 Interim summary

Table 11 summarizes the results of the models using the different groups of predictors. As is

Type of predictors	Measure	Effects found	Explanation		
Categorical	Critical verb RTs	Sentence voice	active	>	passive
	Spillover word RTs	Verb type × Sentence voice	objexp active	>	subjexp active
			subjexp passive	>	objexp passive
			subjexp passive	>	subjexp active
			objexp active	>	objexp passive
	Question accuracy	Sentence voice	passive	<	active
		Verb type × Sentence voice	objexp active	<	subjexp active
			subjexp passive	<	objexp passive
			subjexp passive	<	subjexp active
			objexp active	<	objexp passive
Frequency	Spillover word RTs	Sentence voice frequency	higher frequency	↔	lower RT
	Question accuracy	Sentence voice frequency	higher frequency	↔	higher accuracy
Both	Critical verb RTs	Sentence voice	passive	>	active
	Spillover word RTs	Verb type × Sentence voice	objexp active	>	subjexp active
			subjexp passive	$>^{m}$	objexp passive
			objexp active	>	objexp passive
	Question accuracy	Sentence voice	passive	<	active
		Question voice	active	<	passive
		Verb type × Sentence voice	objexp active	<	subjexp active
			subjexp passive	<	objexp passive
			subjexp passive	<	subjexp active
			objexp active	<	objexp passive

Table 11: Summary of model results (Experiment 1). < indicates lower RTs and accuracy; > indicates the opposite. ↔ indicates a relationship between a continuous predictor (on the left) and a dependent variable (on the right).

clear from table 11, the models fit using both kinds of predictors mostly found the same effects found by the models that did not use frequency as a predictor. The exception to this was that for the spillover word, the models with both kinds of predictors found no effect of Sentence voice for subject experiencer verbs, while the model without frequency did show this effect. In contrast, while the models using only frequency as a predictor found frequency-related effects on the spillover word RTs and question accuracy, the models with both kinds of predictors found no effects of frequency.

Thus, we seem to have two possible explanations of our effects, which are compatible with different kinds of models. Models using the categorical predictors showed effects related to those predictors, while models using only frequency find effects of frequency. However, it is not obvious whether the models using grammatical predictors better explain the results than the models using only frequency. For this reason, we conducted comparisons of the different kinds of models, as detailed in

section C.1.4.

C.1.4 Model comparisons

We compared the models presented in the preceding sections to determine the relative utility of frequency compared to our categorical predictors for crossed and nested models. In addition, we compared models fit using both frequency and categorical predictors. In addition to these models and the models we discussed before, which all used the maximal random effects structure, we also fit models that systematically added categorical or frequency-based fixed effect(s) but not the corresponding random effect(s). This was done in order to examine the predictive utility of adding a single type of fixed effect(s) at a time without considering the possible ramifications of simultaneously making the random effects structure more complex.

We note here that the most crucial comparisons are between the frequency-only and Sentence voice × Verb type models. This is important because these model structures are not nested within each other. Thus, even though the Sentence voice × Verb type models have more predictors than the frequency-only models, it is not necessarily the case that the model with more predictors will explain more variance than the model with fewer predictors. This is in contrast to comparisons of models where predictors are added to another model, where the model with a superset of predictors will always account for more variance than the model with the subset of predictors. Even in our additive comparisons, where this is true, we consider not which model explains more variance, but whether the increased variance explained merits the inclusion of the additional predictor according to statistical criteria.

Our comparisons examine the (log) Bayes Factor (Nicenboim et al. 2022 ch. 15): the natural log of the ratio of the marginal likelihoods of some dataset for two models, defined simply as follows:

$$LBF_{12} = \ln\left(\frac{P(y|\mathcal{M}_1)}{P(y|\mathcal{M}_2)}\right) \tag{10}$$

where y is some dataset, $\mathcal{M}$ is a model, $P(y|\mathcal{M})$ is the posterior probability assigned to y by $\mathcal{M}$, and LBF_{12} the log Bayes factor that compares $\mathcal{M}_1$ to $\mathcal{M}_2$. If both models assign equal probability to y, $LBF_{12} = 0$; $LBF_{12} > 0$ means that $\mathcal{M}_1$ better accounts for y than $\mathcal{M}_2$ (it assigns higher probability to it). Conversely, $LBF_{12} < 0$ means that $\mathcal{M}_2$ better accounts for y than $\mathcal{M}_1$. We estimate $P(y|\mathcal{M})$ using bridge sampling, implemented in R's `bridgesampling` library (Gronau et al. 2020). To account for variability in estimates, we run the bridge sampler 100 times per model, and obtain log Bayes factors based on the medians of these 100 estimates using the `brms` package (Bürkner 2017, 2018, 2021).

Nicenboim et al. (2022) offer the "rules of thumb" for interpreting Bayes factors (from Jeffreys 1939) presented in table 12.[9] While these are not intended as categorical rules, we present the kind of evidence they indicate in the following tables solely for ease of interpretation. We present the "total ordering" implied by these rules of thumb at the bottom of each table where applicable.

In the following tables, we abbreviate our fixed effects as follows: SV (Sentence voice), QV (Question voice), VT (Verb type), SV ∈ SE (Sentence voice (subjexp)), SV ∈ OE (Sentence Voice (objexp)), VT ∈ A (Verb type (active)), VT ∈ P (Verb type (passive)), SFq (Sentence voice log odds), QFq (Question voice log odds). In addition, we refer to comparisons of models with the full random effects

[9] Nicenboim et al. (2022) present this table for the non-log version of Bayes factors; we converted their numbers to the natural log space since we present log Bayes factors. We present log Bayes factors because some of the Bayes factors we found were large enough that exponentiating them to obtain the raw values resulted in numbers too large for the 32-bit float format.

LBF_{12}	Interpretation
> 4.6	Extreme evidence for $\mathcal{M}_1$
> 3.4	Very strong evidence for $\mathcal{M}_1$
> 2.3	Strong evidence for $\mathcal{M}_1$
> 1.1	Moderate evidence for $\mathcal{M}_1$
> 0	Anecdotal evidence for $\mathcal{M}_1$
0	No evidence
< 0	Anecdotal evidence for $\mathcal{M}_2$
< -1.1	Moderate evidence for $\mathcal{M}_2$
< -2.3	Strong evidence for $\mathcal{M}_2$
< -3.4	Very strong evidence for $\mathcal{M}_2$
< -4.6	Extreme evidence for $\mathcal{M}_2$

Table 12: Rules of thumb for interpreting log Bayes factors from Jeffreys (1939), via Nicenboim et al. (2022)

structures as "predictive" comparisons, and comparisons of models with non-maximal random effects structures as "additive" comparisons. By "×," we refer to models that contain all main effects of the relevant factors as well as all interactions (i.e., "SV × VT" indicates that the model contains main effects of both Sentence voice and Verb type, as well as their interaction). We round the log Bayes factors to two decimal places for presentation.

For perspicuity, we notate factors that were added to the fixed effects structure but not to the random effects structure in italics. These models with a non-maximal random effects structure were used only for additive comparisons, to see whether adding a type of fixed effect(s) to another model improved it. Upright text means an effect was included in both the fixed and random effects for subjects and items. Models using the full random effects structures justified by the fixed effects were compared to determine which model best predicted the data.

Critical verb

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SFq}})}$	6363.87	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \text{SFq}})}$	6386.86	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \text{SFq}})}$	23.00	Extreme evidence for SFq
Total ordering: SV × VT > SFq > SV × VT + SFq		

Table 13: Critical verb RTs predictive crossed model log Bayes factor comparisons (Experiment 1)

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P}})}{P(y\|\mathcal{M}_{\text{SFq}})}$	6363.78	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P(y\|\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P}})}{P(y\|\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P + SFq}})}$	6387.97	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P + SFq}})}$	24.19	Extreme evidence for SFq
Total ordering: SV + VT ∈ A + VT ∈ P > SFq > SV + VT ∈ A + VT ∈ P + SFq		

Table 14: Critical verb RTs predictive Verb type nested in Sentence voice model log Bayes factor comparisons (Experiment 1)

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE}})}{P(y\|\mathcal{M}_{\text{SFq}})}$	6364.71	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P(y\|\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE}})}{P(y\|\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}})}$	6384.60	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}})}$	19.90	Extreme evidence for SFq
Total ordering: VT + SV ∈ SE + SV ∈ OE > SFq > VT + SV ∈ SE + SV ∈ OE + SFq		

Table 15: Critical verb RTs predictive Sentence voice nested in Verb type model log Bayes factor comparisons (Experiment 1)

Predictive comparisons In general, across the predictive comparisons for the critical verb, we found that the models not including frequency were better predictors of the data than the models with only frequency, and these in turn were better predictors than the models that combined both types of predictors.

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \textit{SFq}})}$	6358.54	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{SV} \times \textit{VT} + \text{SFq}})}$	5.29	Extreme evidence for SFq
$\frac{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P}})}{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \textit{SFq}})}$	6360.45	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{SV} + \textit{VT} \in \textit{A} + \textit{VT} \in \textit{P} + \text{SFq}})}$	5.45	Extreme evidence for SFq
$\frac{P(y\|\mathcal{M}_{\text{VT} + \text{SV} \in \text{SE} + \text{SV} \in \text{OE}})}{P(y\|\mathcal{M}_{\text{VT} + \text{SV} \in \text{SE} + \text{SV} \in \text{OE} + \textit{SFq}})}$	6360.86	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{VT} + \textit{SV} \in \textit{SE} + \textit{SV} \in \textit{OE} + \text{SFq}})}$	4.83	Extreme evidence for SFq

Table 16: Critical verb RTs additive model log Bayes factor comparisons (Experiment 1)

Additive comparisons The pattern of results is similar to what we found for the predictive comparisons for the additive comparisons. Models without frequency fared better than models with frequency added to the fixed effects. In turn, models with frequency as the only predictor fared better than models that added categorical predictors to frequency.

Spillover word

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SFq}})}$	6382.18	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \text{SFq}})}$	6380.18	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \text{SFq}})}$	−2.00	Moderate evidence for SV × VT + SFq
Total ordering: SV × VT > SV × VT + SFq > SFq		

Table 17: Spillover word RTs predictive crossed model log Bayes factor comparisons (Experiment 1)

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P}})}{P(y\|\mathcal{M}_{\text{SFq}})}$	6379.12	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P}})}{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \text{SFq}})}$	6380.12	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \text{SFq}})}$	1.00	Anecdotal evidence for SFq
Total ordering: SV + VT ∈ A + VT ∈ P > SFq > SV + VT ∈ A + VT ∈ P + SFq		

Table 18: Spillover word RTs predictive Verb type nested in Sentence voice model log Bayes factor comparisons (Experiment 1)

Comparison	LBF	Interpretation
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}$	6381.90	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}$	6378.60	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}$	−3.30	Strong evidence for VT + SV ∈ SE + SV ∈ OE + SFq
Total ordering: VT + SV ∈ SE + SV ∈ OE > VT + SV ∈ SE + SV ∈ OE + SFq > SFq		

Table 19: Spillover word RTs predictive Sentence voice nested in Verb type model log Bayes factor comparisons (Experiment 1)

Predictive comparisons Like for the critical verb, we found that the model without frequency consistently served as a better predictor of the data for the spillover word. Unlike for the critical verb, we found in some cases, there was some evidence that models with categorical predictors and frequency were better predictors of the data than models using frequency alone (the crossed and Sentence voice nested models). For the Verb type nested model, we found that the model with frequency alone was a better predictor, though only at an "anecdotal" level.

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \textit{SFq}})}$	6356.52	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{SV} \times \textit{VT} + \text{SFq}})}$	−2.29	Moderate evidence for $SV \times VT$ + SFq
$\frac{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P}})}{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \textit{SFq}})}$	6357.25	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{SV} + \textit{VT} \in \textit{A} + \textit{VT} \in \textit{P} + \text{SFq}})}$	−2.42	Strong evidence for $SV + VT \in A + VT \in P$ + SFq
$\frac{P(y\|\mathcal{M}_{\text{VT} + \text{SV} \in \text{SE} + \text{SV} \in \text{OE}})}{P(y\|\mathcal{M}_{\text{VT} + \text{SV} \in \text{SE} + \text{SV} \in \text{OE} + \textit{SFq}})}$	6358.12	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{VT} + \textit{SV} \in \textit{SE} + \textit{SV} \in \textit{OE} + \text{SFq}})}$	−0.24	Anecdotal evidence for $VT + SV \in SE + SV \in OE$ + SFq

Table 20: Spillover word RTs additive model log Bayes factor comparisons (Experiment 1)

Additive comparisons In the additive comparisons for the spillover word, we found that models using only categorical predictors fared better than models that added frequency to the categorical predictors. In addition, models adding categorical predictors to frequency fared better, though for the model adding the Verb type nested categorical predictors to frequency, this evidence was not substantial.

Question responses

Comparison	LBF	Interpretation
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV} \times \text{VT} \times \text{QV}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}$	6390.19	Extreme evidence for SV × VT × QV
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV} \times \text{VT} \times \text{QV}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV} \times \text{VT} \times \text{QV} + \text{SFq}}\right)}$	6377.19	Extreme evidence for SV × VT × QV
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV} \times \text{VT} \times \text{QV} + \text{SFq}}\right)}$	−13.00	Extreme evidence for SV × VT × QV + SFq
Total ordering: SV × VT × QV > SV × VT × QV + SFq > SFq		

Table 21: Question accuracy predictive crossed model log Bayes factor comparisons (Experiment 1)

Comparison	LBF	Interpretation
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}$	6391.23	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \text{SFq}}\right)}$	6376.92	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \text{SFq}}\right)}$	−14.31	Extreme evidence for SV + VT ∈ A + VT ∈ P + SFq
Total ordering: SV + VT ∈ A + VT ∈ P > SV + VT ∈ A + VT ∈ P + SFq > SFq		

Table 22: Question accuracy predictive Verb type nested in Sentence voice model log Bayes factor comparisons (Experiment 1)

Comparison	LBF	Interpretation
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}$	6391.83	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}$	6376.85	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}$	−14.98	Extreme evidence for VT + SV ∈ SE + SV ∈ OE + SFq
Total ordering: VT + SV ∈ SE + SV ∈ OE > VT + SV ∈ SE + SV ∈ OE + SFq > SFq		

Table 23: Question accuracy predictive Sentence voice nested in Verb type model log Bayes factor comparisons (Experiment 1)

Predictive comparisons For the predictive comparisons of the question accuracy models, we found that the models without frequency fared best, followed by the models with both categorical predictors and frequency, followed by the frequency only models.

Comparison	LBF	Interpretation
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV × VT × QV}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV × VT × QV + }\mathit{SFq}}\right)}$	6358.08	Extreme evidence for SV × VT × QV
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\mathit{SV} \times \mathit{VT} \times \mathit{QV} + \text{SFq}}\right)}$	−21.54	Extreme evidence for $SV \times VT \times QV$ + SFq
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + }\mathit{SFq}}\right)}$	6356.41	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\mathit{VT + SV ∈ SE + SV ∈ OE} + \text{SFq}}\right)}$	−21.59	Extreme evidence for $VT + SV ∈ SE + SV ∈ OE$ + SFq
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P + }\mathit{SFq}}\right)}$	6357.35	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\mathit{SV + VT ∈ A + VT ∈ P} + \text{SFq}}\right)}$	−21.95	Extreme evidence for $SV + VT ∈ A + VT ∈ P$ + SFq

Table 24: Question accuracy additive model log Bayes factor comparisons (Experiment 1)

Additive comparisons Across the different additive model comparisons for question accuracy, we found that models with only categorical predictors fared better than those that added frequency as a predictor. In addition, we found that models adding categorical predictors to frequency consistently fared better models using frequency alone.

C.2 Experiment 2

C.2.1 Voice frequency model

As with experiment 1, models fit using both categorical and frequency predictors showed a similar pattern of results to the models fit using only categorical predictors. We leave a full presentation of these results to Appendix C.2.2, but present the effects found in the interim summary (section C.2.3) in brief.

In experiment 1, models using frequency only found effects, but models using only categorical predictors generally fared better than models using both frequency and categorical predictors, which in turn fared better than models that used only frequency as a predictor. However, it is possible that frequency may have different effects on different kinds of measures. Although we found evidence that our categorical predictors were better at accounting for variability in reading times and response accuracy, that does not mean that frequency might not account for the patterns in grammaticality judgments better than our categorical predictors. In particular, the more exposure a comprehender has to a particular verb in a particular voice, the less marked they might perceive that configuration to be relative to less frequent combinations. For this reason, we also fit models using only frequency and models combining both kinds of predictors for the results of experiment 2, and compared these models to determine which were better supported using a Bayes factor analysis (see section C.1.4 for explanation).

Figure 8 shows the regression of the mean "grammatical" responses on the log odds of the verb occurring in the voice it did in the sentence. A frequency-based approach to grammaticality judgments predicts a line with a positive slope, since more frequent exposure to a verb in a particular configuration should lead to it being considered less marked and thereby more often grammatical. As before, we used models that took into account measurement error for the frequency predictor in our Bayesian regressions, though we omit error bars from the x-axis in figure 8 for visual clarity.

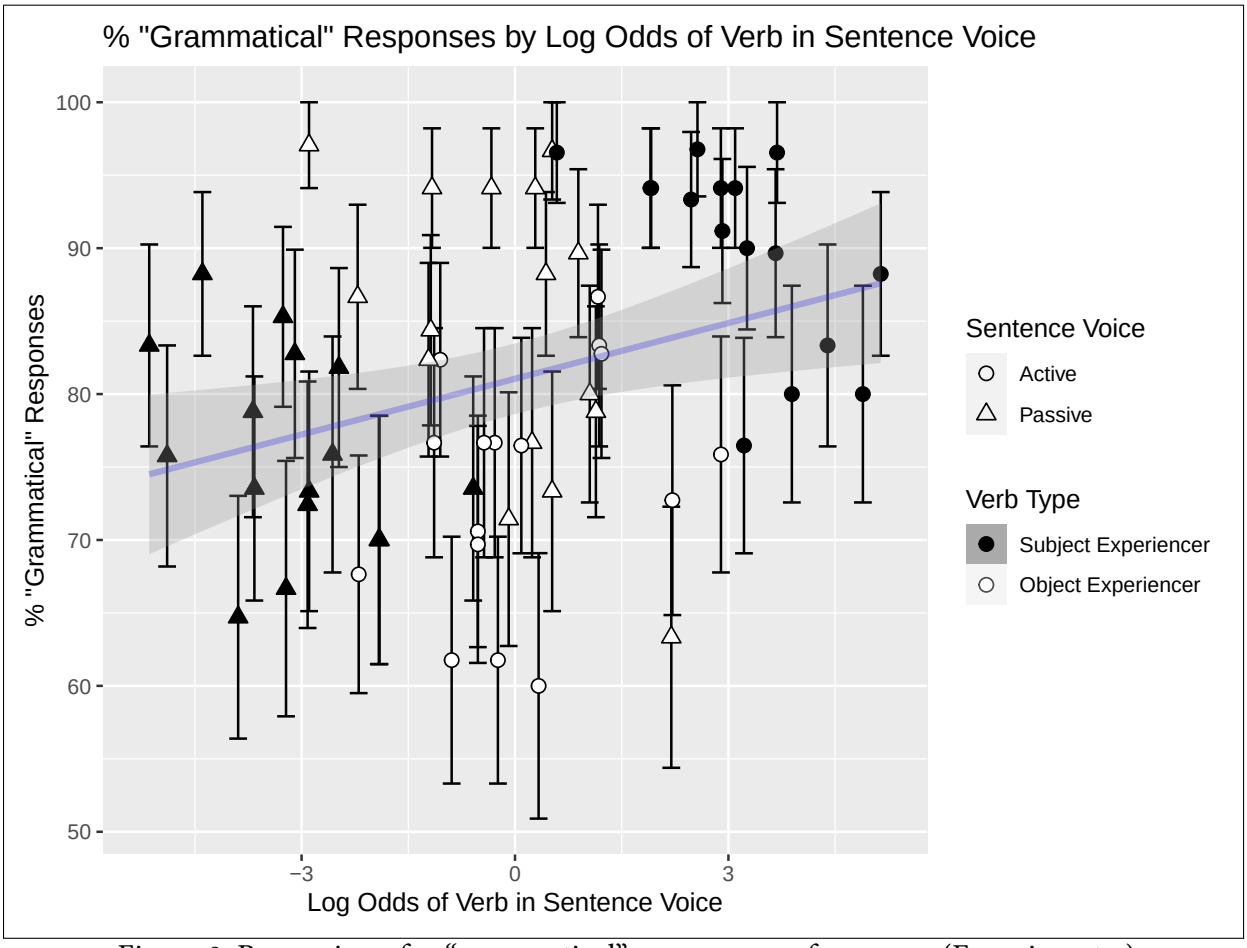


Figure 8: Regression of % "grammatical" responses on frequency (Experiment 2)

Table 25 shows the results of the logistic regression of Response on frequency. The model found no effect.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	1.89	0.18	1.54	2.26
Sentence voice log odds	0.07	0.04	−0.02	0.14

Table 25: Frequency model for Response (Experiment 2)

C.2.2 Experiment 2: Voice, Verb type, and Voice frequency model

As in experiment 1, we fit models combining both categorical and frequency-based predictors to determine the potential relevance of both sources of variance, as well as to set us up for a comparison of models.

Table 26 reports the results of the model crossing our categorical predictors and adding frequency.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	2.09	0.20	1.71	2.50
Verb type	−0.41	0.19	−0.79	−0.05
Sentence voice	0.28	0.22	−0.17	0.72
Verb type × Sentence voice	−2.03	0.45	−2.91	−1.16
Sentence voice log odds	−0.02	0.04	−0.11	0.07

Table 26: Crossed model for Response with frequency (Experiment 2)

Like the model using only categorical predictors, the model combining the categorical predictors with frequency found a main effect of Verb type (95% CI: $[-0.79, -0.05]$) and an interaction of Verb type and Sentence voice (95% CI: $[-2.91, -1.16]$) Like the frequency only model, no effect of frequency was found (95% CI: $[-0.11, 0.07]$)

To examine the interaction in this model, we fit models nesting Verb type in Sentence voice and nesting Sentence voice in Verb type. As in the models that used only categorical predictors, the nested models with frequency showed all four effects. Model results are reported in tables 27 and 28.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Verb type (active)	−1.60	0.35	−2.30	−0.93
Verb type (passive)	0.72	0.27	0.18	1.24

Table 27: Verb type nested in Sentence voice model for Response with frequency (Experiment 2)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI
Sentence voice (subjexp)	1.41	0.37	0.69	2.14
Sentence voice (objexp)	−0.85	0.28	−1.41	−0.30

Table 28: Sentence voice nested in Verb type model for Response with frequency (Experiment 2)

C.2.3 Interim summary

Table 29 summarizes the results of the models of Response for experiment 2 using the different groups of predictors. The models fit using both kinds of predictors found at least the effects also found by the models that did not use frequency as a predictor, with the additional finding of a main effect of Sentence voice that was not found in the model without frequency. In contrast, while the model fit using only frequency found an effect of frequency, no such effect was found in the model that included the categorical predictors as well.

We now turn to a comparison of these different models to determine which factor(s) best predict the response data.

Type of predictors	Effects found	Explanation
Categorical	Verb type	objexp < subjexp
	Verb type × Sentence voice	objexp active < subjexp active subjexp passive < objexp passive subjexp passive < subjexp active objexp active < objexp passive
Both	Verb type	objexp < subjexp
	Sentence voice	passive $<^{m}$ active
	Verb type × Sentence voice	objexp active < subjexp active subjexp passive < objexp passive subjexp passive < subjexp active objexp active < objexp passive

Table 29: Summary of model results (Experiment 2). < indicates that the condition on the left was judged grammatical less often than the condition on the right. ↔ indicates a relationship between a continuous predictor (on the left) and the proportion of "grammatical" responses.

C.2.4 Model comparisons

For experiment 2, we used the same kinds of models for comparison purposes that we did for experiment 1. As with experiment 1, we conducted comparisons using log Bayes factors (rounded to two decimal places in our tables).

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SFq}})}$	5439.00	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \text{SFq}})}$	5420.87	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \text{SFq}})}$	−18.13	Extreme evidence for SV × VT + SFq
Total ordering: SV × VT > SV × VT + SFq > SFq		

Table 30: Response predictive crossed model log Bayes factor comparisons (Experiment 2)

Comparison	LBF	Interpretation
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}$	5440.33	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P + SFq}}\right)}$	5421.78	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SV + VT ∈ A + VT ∈ P + SFq}}\right)}$	−18.55	Extreme evidence for SV + VT ∈ A + VT ∈ P + SFq
Total ordering: SV + VT ∈ A + VT ∈ P > SV + VT ∈ A + VT ∈ P + SFq > SFq		

Table 31: Response predictive Verb type nested in Sentence voice model log Bayes factor comparisons (Experiment 2)

Comparison	LBF	Interpretation
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}$	5440.17	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}$	5421.70	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P\left(y\middle\vert\mathcal{M}_{\text{SFq}}\right)}{P\left(y\middle\vert\mathcal{M}_{\text{VT + SV ∈ SE + SV ∈ OE + SFq}}\right)}$	−18.47	Extreme evidence for VT + SV ∈ SE + SV ∈ OE + SFq
Total ordering: VT + SV ∈ SE + SV ∈ OE > VT + SV ∈ SE + SV ∈ OE + SFq > SFq		

Table 32: Response predictive Sentence voice nested in Verb type model log Bayes factor comparisons (Experiment 2)

Predictive comparisons In experiment 2, we found that that the models using only categorical predictors were consistently better predictors than those using both frequency and categorical predictors, which in turn fared better than the models using only frequency. This general pattern matched the results of the comparisons from experiment 1.

Comparison	LBF	Interpretation
$\frac{P(y\|\mathcal{M}_{\text{SV} \times \text{VT}})}{P(y\|\mathcal{M}_{\text{SV} \times \text{VT} + \textit{SFq}})}$	5391.21	Extreme evidence for SV × VT
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{SV} \times \textit{VT} + \text{SFq}})}$	−19.27	Extreme evidence for *SV* × *VT* + SFq
$\frac{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P}})}{P(y\|\mathcal{M}_{\text{SV} + \text{VT} \in \text{A} + \text{VT} \in \text{P} + \textit{SFq}})}$	5391.07	Extreme evidence for SV + VT ∈ A + VT ∈ P
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{SV} + \textit{VT} \in \textit{A} + \textit{VT} \in \textit{P} + \text{SFq}})}$	−20.37	Extreme evidence for *SV* + *VT* ∈ *A* + *VT* ∈ *P* + SFq
$\frac{P(y\|\mathcal{M}_{\text{VT} + \text{SV} \in \text{SE} + \text{SV} \in \text{OE}})}{P(y\|\mathcal{M}_{\text{VT} + \text{SV} \in \text{SE} + \text{SV} \in \text{OE} + \textit{SFq}})}$	5391.18	Extreme evidence for VT + SV ∈ SE + SV ∈ OE
$\frac{P(y\|\mathcal{M}_{\text{SFq}})}{P(y\|\mathcal{M}_{\textit{VT} + \textit{SV} \in \textit{SE} + \textit{SV} \in \textit{OE} + \text{SFq}})}$	−19.99	Extreme evidence for VT + SV ∈ SE + SV ∈ OE + *SFq*

Table 33: Response additive model log Bayes factor comparisons (Experiment 2)

Additive comparisons The additive comparisons found that the models including only categorical predictors fared better than those that added frequency to the categorical predictors. In addition, the model using only frequency fared worse than the models adding categorical predictors to frequency.

C.3 Discussion

In this appendix, we compared models that used categorical predictors implicated in argument structure linking constraints (Verb type and Sentence/Question voice) to models using frequency (the log odds of a verb occurring in the voice it did in a particular sentence). Adding frequency produced no notable improvement in predictive utility, while adding grammatical predictors consistently improved models' performance. However, the best models were those that used only categorical predictors without frequency. The fact that the models without frequency as a predictor received the most support is most consistent with the view that frequency does not explain the results we found.

We do not conclude that frequency plays no role in comprehension or processing—instead, we simply conclude that it was not the strongest predictor of our results, and we suggest that the same may be true for similar sorts of alignment effects that have been found in the literature (cf. Huang et al. 2023; Huber et al. 2023). Though a perennial concern with frequency-based measures is that estimates depend entirely on the coverage of the selected corpus, our estimates are overall remarkably similar to those reported in Gennari & MacDonald (2009) which were collected on the basis of different structures, verbs, and corpora. While this does not guarantee our measure is reflective

of the true distribution of participants' experience, this consistency is encouraging. There is a trend in more recent work (Huang et al. 2023; Huber et al. 2023; Smith & Levy 2013) to use surprisal values obtained from computational language models as a frequency-based predictor of behavioral measures in psycholinguistic experiments, rather than counts obtained directly via corpus studies; we leave the question of whether using this alternative proxy for participants' prior experience with particular voice and verb type combinations would affect our results for future work.

More generally, even if changing to an alternative frequency-based predictor did affect the results of our model comparisons, it would not obviate the need to address important questions about how frequency distributions arise, and how they influence comprehension processes. Work that assumes Surprisal Theory often sidesteps this issue, and treats distributional frequencies as given, without attempting to explain their source. Notably, Gennari & MacDonald (2009)'s Production-Distribution-Comprehension (PDC) approach is an approach that treats frequency as an important predictor, but does attempt to address this question. PDC treats distributional frequencies as arising from production preferences that are imposed by non-linguistic cognitive constraints. But it is likely that there are additional sources of distributional patterns. In particular, Argaman & Pearlmutter (2002) propose treating lexical semantic factors as partly responsible for distributional frequencies. They show that uses of verbs and nouns that share roots with the same meaning (e.g., *accuse*, *accusation*) have correlated frequencies of use, which argues against a "random variations" approach to frequency, that nevertheless differs from the PDC and grammatical constraint-based frameworks. The idea that lexical semantics influences frequency biases is consistent with our approach. In particular, we expect frequency to reflect lexical semantic properties like thematic relations and markedness configurations that arise from them linking up with particular syntactic environments. However, what determines frequency overall will be a combination of these lexical semantic factors and other factors, including (non-linguistic) cognitive factors, as well as pragmatic and discourse factors. Thus, we expect a relationship between frequency and (un)marked structures, but an indirect one at best. This could account for why we found weaker evidence for effects of frequency in our experiments than might be predicted by an approach which posits a single underlying source for production and processing effects related to an experiencer-subject preference, as well as why the results of our corpus study did not perfectly mirror what we would expect if producers uniformly favored aligned configurations. The frequencies observed in a corpus necessarily reflect the interaction of more factors than do the intentionally factitious contexts used in an experimental setting.

An additional factor that might play a role is the overall frequency of the kinds of sentences we used as stimuli. It could be the case that comprehenders rely more on frequency when processing highly frequent structures than when processing relatively infrequent structures (though this is not typically assumed by Surprisal Theory, where the most notable predictions are those for less frequently encountered structures). Possibly, the structures used our stimuli (at whatever level they might be characterized) were sufficiently infrequent that comprehenders lacked enough experience with them to rely on frequency information heavily, and instead had to rely on grammatical markedness instead. In order to evaluate this claim, however, it is important to determine what kinds of structures frequency is recorded for. For instance, in our stimuli, the relevant measure might be the overall frequency of RCs, ORCs, prepositional object RCs, RCs with experiencer verbs, RCs with experiencer verbs with two animate arguments, presentational sentences, etc. But even if something like this does turn out to be the case, it only means that our conclusions show that when frequency of particular sentence types is low overall, participants' processing and comprehension is affected by alignment in the way predicted by the experiencer-subject preference, and contra the predictions

of an experiencer-first preference, at least in English.

D. Experiment 1 analyses with filtered RTs

As we noted in the text, we did not use any criteria to filter out reading time data in experiment 1. However, in addition to the analysis we included in the text, we also conducted the same set of analyses when filtering out RTs < 100 ms or > 3000 ms. These led to a similar pattern of results as what we found in the text. We summarize the results in table 1. (And of course, filtering reading times did not impact the Question accuracy models, which were not based on reading time data.)

Type of predictors	Measure	Effects found	Explanation
Categorical	Critical verb RTs	Sentence voice	(objexp) active > (objexp) passive[10]
	Spillover word RTs	Verb type × Sentence voice	objexp active > subjexp active objexp active > objexp passive
Both	Critical verb RTs	Sentence voice	objexp active > objexp passive
	Spillover word RTs	Verb type × Sentence voice	objexp active > subjexp active objexp active > objexp passive

Table 1: Summary of model results with filtered RTs (Experiment 1)

We note that the full crossover interaction was not found when RTs were filtered. In particular, the effect of Verb type in passive sentences for the spillover word was not found, nor was the effect of Sentence voice for subject experiencer verbs. However, the models still found that subject experiencer active sentences were read faster than object experiencer active sentences, and that object experiencer passive sentences were read faster than object experiencer active sentences.

We believe there is good reason to be skeptical of over-interpreting these differences caused by filtering out some of the RTs, particularly when considering the pattern of exactly which RTs were filtered out. Table 2 shows the number of observations removed in each condition for the critical verb and spillover word. Prior to filtering, the number of points in each cell was 608 (2,432 observations per word total). The number of observations < 100 ms and > 3000 ms that were excluded are included in parentheses after the total number of remaining observations, separated by a semicolon.

Word	Verb type	Sentence voice	
		Active	Passive
Critical verb	SubjExp	599 (3; 6)	599 (4; 5)
	ObjExp	588 (4; 16)	605 (0; 3)
Spillover word	SubjExp	599 (5; 4)	570 (8; 30)
	ObjExp	580 (5; 23)	599 (1; 8)

Table 2: Observations per condition after filtering RTs < 100 ms or > 3000 ms (Experiment 1)

[10] There was a main effect of Sentence voice in the crossed model, but no interaction. However, in the model nesting Verb type within Sentence voice, the effect of Sentence voice was only found for object experiencer verbs.

There are a similar number of exclusions across conditions and measures for RTs < 100 ms, but the highest numbers of exclusions (8 subject experiencer exclusions in passive sentences) represent data that is disfavorable to our general findings, as it would overestimate the ease of processing subject experiencer verbs in passive sentences.

More striking, however, is the pattern of exclusions for RTs > 3000 ms. In particular, there is a clear relationship between the number of points > 3000 ms excluded and the condition. The highest number of exclusions happen in passive subject experiencer sentences and active object experiencer conditions, especially for the spillover word. These are precisely the conditions where participants showed the most difficulty with comprehension questions. As such, it is unsurprising that RTs for the verb and post-verbal material in these conditions would be longer, if participants were taking extra time to process the sentence in an effort to answer comprehension questions more accurately. For this reason, we believe that filtering RTs based on these criteria is not the ideal way of examining our results—doing so seems to exclude data that is relevant to answering the question of which kinds of sentences language users find particularly difficult to process and comprehend.

In addition, while the reading time results from experiment 1 may not be as clear when filtering out RTs based on these criteria, we note that the question accuracy data from experiment 1 and speeded grammaticality judgment data from experiment 2 provide strong evidence on their own for the effects we found, at the very least in the domain of comprehension if not processing. Thus, we do not believe filtering or not filtering RTs according to the criteria we described makes a substantial difference for how we interpreted our results.

E. Suspended eye-tracking study

In addition to the SPR and speeded grammaticality judgment studies, we also had begun collecting data for an eye-tracking while reading study in Fall 2019 and Spring 2020. However, due to safety protocols put into place because of the COVID-19 pandemic and subsequent unrelated events, data collection has been suspended indefinitely. We collected data from only 15 participants, who were undergraduates at the University of Massachusetts Amherst and received course credit for their participation. Otherwise, they had to meet the same criteria as described in experiment 1 (with the exception that there was no pre-experiment questionnaire). All participants gave informed consent.

The same items and fillers were used as in experiment 1. However, we note that 10 of our 15 participants saw two items with typos (one with a missing period at the end of the sentence, and another using the typo'd word *businessman**n***).

Before beginning the experiment, the eye-tracker was calibrated using nine points at the center of the screen, the middle of each side of the screen, and each corner. At the beginning of each trial, participants saw a black gaze-change fixation box on the left side of the screen. Upon looking at it, a sentence would be displayed in its entirety, vertically centered on the screen in a fixed-width font (Monaco). Participants read the sentence for as long as they wanted, and pushed a button on a video game controller to proceed to a question. Participants responded to questions by pressing the left or right button on the controller, after which the next trial would begin. The experimenters offered participants breaks approximately ⅓ and ⅔s of the way through the experiment. The experimenter was separated from the participants by a hanging sheet, but observed their recorded eye movements in real-time on a nearby monitor. If tracking became decalibrated, the experimenter would stop the experiment and recalibrate before proceeding. The experiment took about 40 minutes total.

We used Python's `sideeye` library (`github.com/amnda-d/sideeye`) along with a script avail-

able at the first author's GitHub (github.com/mawilson1234/prASC), to format data from the ASCs collected by the EyeLink eye-tracker for analysis. Regions of analysis are shown for a sample item in (11).

(11) a. That's the actor/ that the director/ particularly/ appreciates/ because of/ his dramatic emotional bearing.
(active, subject experiencer)

b. That's the actor/ that the director/ particularly/ bothers/ because of/ his dramatic emotional bearing.
(active, object experiencer)

c. That's the director/ that the actor/ is particularly/ appreciated by/ because of his dramatic emotional bearing.
(passive, subject experiencer)

d. That's the director/ that the actor/ is particularly/ bothered by/ because of his dramatic emotional bearing.
(passive, object experiencer)

We have not fit any regressions to these data since we did not reach our desired *N*; nevertheless, we present and briefly discuss plots here as they seem consistent with our other results and offer additional suggestive information about how participants process misaligned sentences.

Figure 9 shows first pass, go past, reread times, and total reading times by region and condition. First pass times refer to the amount of time a participant's gaze stays within a particular region before exiting to the right (i.e., the sum of all fixations made within a region upon entering from the left and before exiting to the right for the first time). Go past times refer to the amount of time a participant's gaze takes to exit a region from the right, starting from the first time it enters from the left. In other words, it includes first pass time along with any time spent looking back to a prior region. Rereading times include the entire amount of time spent reading a region after the first time a participant's gaze exits it to the right (i.e., the total time they spend going back to that region after passing it the first time). Finally, total time measures the entire amount of time a participant spends looking at a particular region.

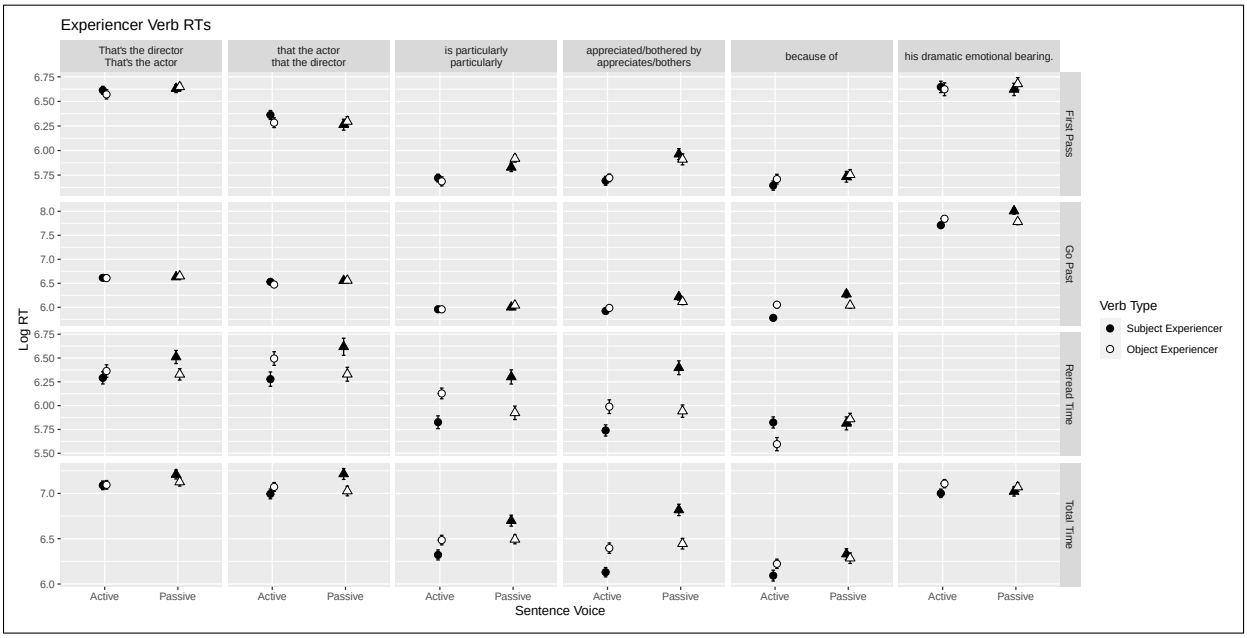


Figure 9: Log RTs for four measures by region, Verb type, and Sentence voice (Eye-tracking Pilot)

While we again note that this discussion is tentative, we notice several patterns that are consistent with the results of experiments 1 and 2, and which provide more insight about how participants might deal with marked configurations. In particular, note that first pass reading times appear to show little evidence of the patterns we found in SPR and speeded grammaticality judgment measures. Thus, when reading more naturally, the effects we found in word-by-word non-cumulative SPR (where regression is impossible) do not appear clearly in this early measure. On the other hand, we begin to see an effect that looks like the crossover interaction we found in experiments 1 and 2 in the go past times for the spillover region (but not before). We can thus suppose that participants read through the beginning of the sentence fairly quickly (as shown by the lack of differences in first pass times and go past times prior to the spillover region), but noticed more difficulty with more marked alignment configurations in the spillover region. The rereading times support this account, as they show the same pattern of difficulty (though more extreme) in all regions prior to the spillover region. Finally, we note that the reread and total times for the matrix subject and embedded subject regions are higher than the total times spent in other regions (with the exception of the final region). This indicates that participants typically resolved the difficulty first encountered in the spillover region by regressing to earlier parts of the sentence more often for the more difficult configurations, and that they spent most of that rereading time focusing on the arguments that they needed to map to the thematic roles of the verb.

Figure 10 is consistent with this account as well. It shows the percentage of first pass regressions in and first pass regressions out by condition and region. Note that the first pass percent regressions out show that in the spillover region, participants made more regressions out in the conditions with the more marked linking configuration compared to the less marked one of the same voice. Furthermore, similar patterns show where participants were regressing to: the interaction effect shows up numerically in the embedded subject, pre-critical, and critical regions as well. This is consistent with the patterns found in the RT plot.

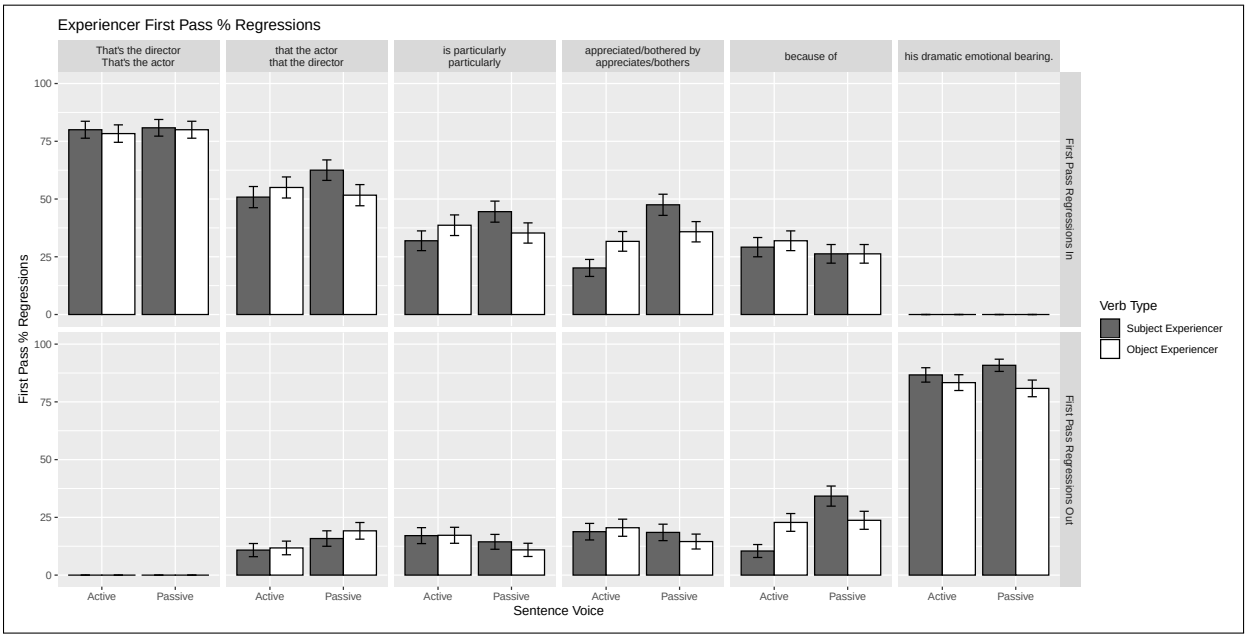


Figure 10: % First pass regressions in/out by region, Verb type, and Sentence voice (Eye-tracking Pilot)

Finally, the question accuracy data from our pilot study shows a similar pattern to what we found in experiment 1. In particular, just as in experiment 1, participants responded correctly less often to active object experiencer verbs and passive subject experiencer verbs than to passive object experiencer verbs and active subject experiencer verbs.

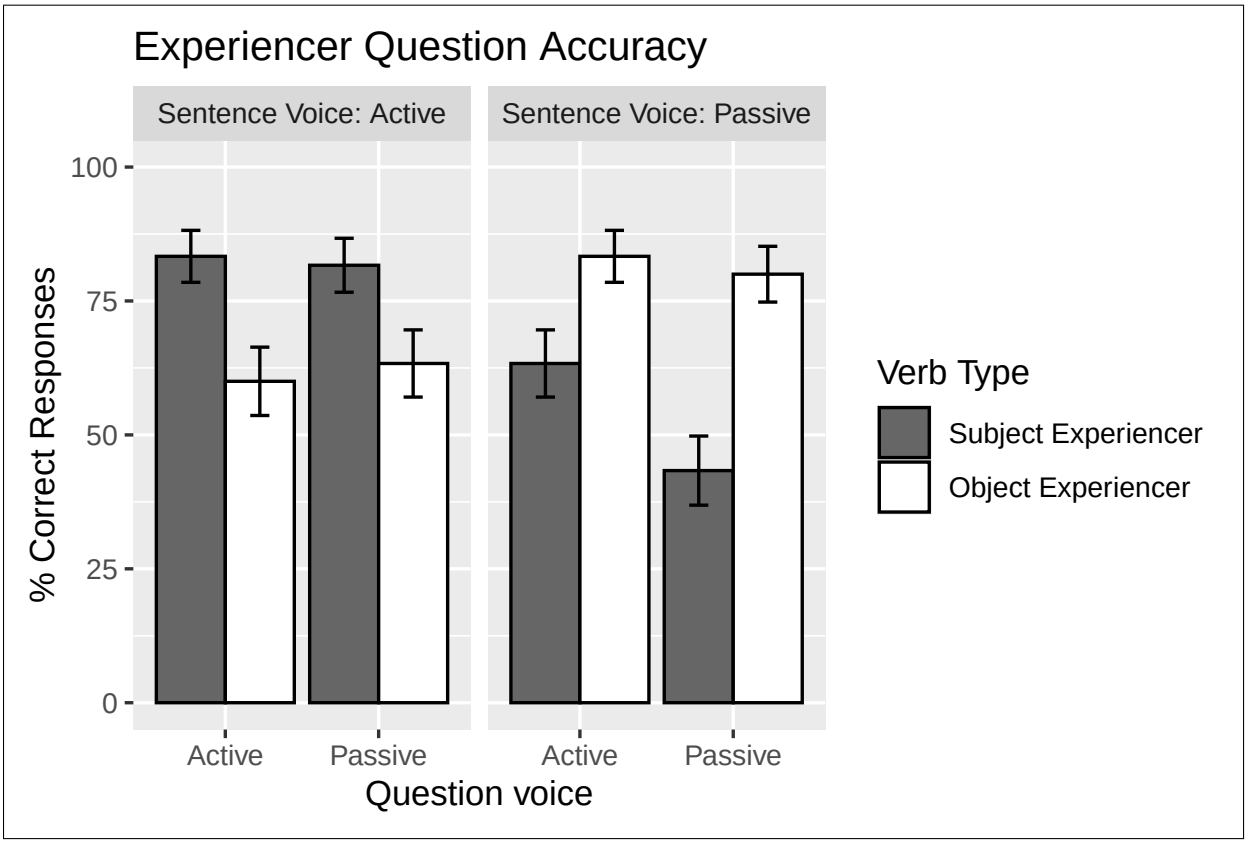


Figure 11: Question accuracy by Verb type, Sentence voice, and Question voice (Eye-tracking pilot)

While we have not conduced a statistical analysis of these data due to not having reached our planned *N*, we believe they are nevertheless suggestive, and certainly compatible with the findings for which we have more experimental evidence presented in the main text. They also suggest that comprehenders may deploy sophisticated strategies in the face of difficulty, as the rereading patterns suggest that participants selectively chose when and where to reread based on what the most difficult parts of the sentence were (as revealed by the question accuracy performance). We hope to resume collection of eye-tracking data when feasible, in order to further examine the additional information the more fine-grained measures this methodology makes available regarding how comprehenders deal with marked configurations.