Alignment between Thematic Roles and Grammatical Functions Facilitates Sentence Processing

Evidence from Experiencer Verbs

Michael Wilson
Yale University

Brian Dillon
University of Massachusetts Amherst

Abstract

How do comprehenders process grammatically marked structures? And how is markedness represented? Previous syntactic, semantic, and psycholinguistic work has studied this question extensively in the domain of experiencer verbs. We consider different psycholinguistic approaches to markedness as it relates to alignment on various grammatical scales, including ones that posit effects of alignment in comprehension (Bornkessel & Schlesewsky 2006), grammatical constraint-based approaches (Wagers et al. 2018), cognitive constraint approaches (e.g., Ferreira 2003), and frequency-based approaches (Gennari & MacDonald 2009).

We evaluate these approaches with two experiments focused on the question of thematic roles' alignment with grammatical functions for experiencer verbs, using self-paced reading with comprehension questions and speeded grammaticality judgments. We find evidence that misaligned configurations are harder to process and understand, and less often considered grammatical compared to aligned configurations. We compare Bayesian models using grammatical predictors only, frequency only, and a combination of the two to determine how frequency influences our results. While there was little predictive difference between models using only grammatical predictors and models using both kinds of predictors, models using frequency only generally fared worse. In models using both kinds of predictors, no evidence for effects of frequency was found. We propose an approach called the Grammar Underlies Production and Processing (plus frequency) (GUPPy) framework where grammatical constraints directly account for alignment effects in production and processing. In addition, our results show the potential importance of directly comparing the distinct causal models of psycholinguistic frameworks.

Keywords

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

Contact information: Michael Wilson, michael.a.wilson@yale.edu, Dow Hall 210A, 370 Temple Street, New Haven, CT, 06511-8957, USA (corresponding author); Brian Dillon, brian@linguist.umass.edu. This research was funded by the University of Massachusetts Amherst.

For valuable discussion and feedback, the authors would like to thank Shota Momma, Lyn Frazier, Shravan Vasishth, Elsi Kaiser, Monica Do, 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 I. Any errors that remain are our own.

1. Introduction

To understand a sentence, comprehenders must link a verb's arguments with their thematic roles. That is, in order to understand (1a), a comprehender must be able to link the subject argument to the agent thematic role, and the object argument to the theme thematic role. In the passive, comprehenders must instead link the theme thematic role to the subject argument, and the agent thematic role to the object of a *by*-phrase, as in (1b).

(1) a. [John]$_{\text{agent}}$ kicked [the ball]$_{\text{theme}}$.
b. [The ball]$_{\text{theme}}$ was kicked by [John]$_{\text{agent}}$.

What factors are relevant to how this linking process occurs in real-time language use? One factor appears to be how canonical the mapping between grammatical function and thematic role is. Non-canonical mappings – as found in passive sentences – famously pose difficulties in language comprehension (e.g., Ferreira 2003) and acquisition (e.g., Nguyen & Pearl 2018, 2021). But the precise nature of this difficulty, and the mechanism by which it impacts real-time comprehension remain under investigation. In this paper we contribute to this body of work by exploring the role of **thematic role-grammatical function alignment** in comprehension with two experiments examining subclasses of experiencer verbs.

Experiencer verbs are of crucial importance in the study of thematic role linking processes. This is because two subclasses of such 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 (2).

(2) a. [John]$_{\text{experiencer}}$ fears [ghosts]$_{\text{theme}}$. (Subject experiencer)
b. [Ghosts]$_{\text{theme}}$ frighten [John]$_{\text{experiencer}}$. (Object experiencer)

Subject experiencer verbs are often considered to present a more canonical mapping between grammatical function and thematic role – for reasons we detail below, we refer to this as a mapping where there is *alignment* between thematic roles and argument roles. In contrast, object experiencer verbs are *misaligned*, presenting a non-canonical mapping.

Our goal is to better understand the role of alignment between grammatical function and thematic role in real-time comprehension. To this end we present two experiments that intentionally control for various factors – including animacy, definiteness, and linear order – in order to focus on the effect of alignment itself. We find that configurations where thematic roles and grammatical functions are aligned are easier to process, easier to comprehend, and rated grammatical more often than misaligned configurations. Our studies thus provide evidence for an important role of thematic alignment that goes beyond the effects of animacy, definiteness, and linear order. We conclude that grammatical hierarchies such as the thematic role hierarchy and the grammatical function hierarchy play an important role in comprehension processes, with our literature review suggesting they play a similar role in production processes. Our results, taken together with this prior literature, suggest that approaches that link effects of alignment to cognitive/linguistic processes that are exclusive to production or comprehension should rather seek a common underlying source, which we suggest resides in grammatical markedness itself. We propose the Grammar Underlies Production and Processing (plus frequency) (GUPPy) framework as an approach compatible with this idea.

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. Then, we review evidence from the psycholinguistic literature that bears on the question of how linking behavior and the alignment of grammatical hierarchies influences processing, before presenting our results.

1.1 Formal syntactic and semantic background

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.[1] They assume that there are universal linking rules that make reference to a universal thematic role hierarchy, which is a partially or totally ordered set of thematic roles. 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, whether these positions are defined in absolute (e.g., Baker 1988) or relative (e.g., Perlmutter & Postal 1984) terms.[2] We can assume the subset of the thematic hierarchy relevant to experiencer verbs is *Experiencer* $\prec$ *Theme* (Experiencer is ranked higher than Theme),[3] and the relevant grammatical function hierarchy is *Subject* $\prec$ *Object*.

Belletti & Rizzi (1988) proposed that object experiencer verbs display misaligned linking properties – in other words, the surface syntax of object experiencer verbs in active transitive sentences does not transparently reflect an optimal linking configuration. While underlyingly, experiencers are always initially projected into a higher hierarchical position than themes, for object experiencer verbs the theme is displaced to become the subject. Though we will not review Belletti & Rizzi (1988)'s evidence for reasons of space, much additional work in the formal syntax and semantics literature has adopted or been inspired by their proposal (Bennis 2004; Bondaruk et al. 2017; Cheung & Larson 2015; Hirsch 2018; Hornstein & Motomura 2002; Landau 2002, 2010; Malaia 2009; Pylkkänen 2000), and it has inspired many psycholinguistic studies (see section 1.2).

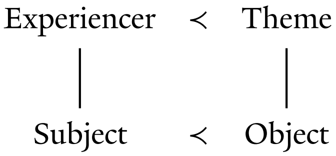
[1] There are proposals that directly argue against Belletti & Rizzi (1988)'s approach, most notably Dowty (1991) and Pesetsky (1995). We avoid discussion of these for reasons of space, but note that the most natural ways in which they make predictions regarding on-line processing behavior are inconsistent with our results.

[2] Although we refer to grammatical function alignment throughout, we are not necessarily committed to this framing. In particular, we believe our proposals are equally compatible with an approach where grammatical functions are substituted for hierarchical (syntactic) positions. We choose to discuss alignment in terms of grammatical functions for perspicuity, since "subject" and "object" are convenient terms for referring to the relevant higher and lower syntactic positions.

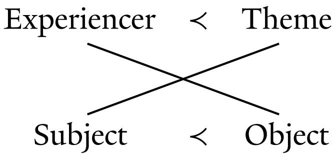
[3] While the precise ranking of thematic roles is debated in different thematic role hierarchy approaches to linking, no approach we are aware of proposes the opposite ranking, *Theme* $\prec$ *Experiencer* (e.g., Baker 1988, 1989; Belletti & Rizzi 1988; Fillmore 1968; Givón 1984; Grimshaw 1990; Jackendoff 1972, 2002; Perlmutter & Postal 1984; Rappaport & Levin 1988; Van Valin 1990).

We can schematize Belletti & Rizzi's account as follows.

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



b. Object experiencer verbs (misaligned):

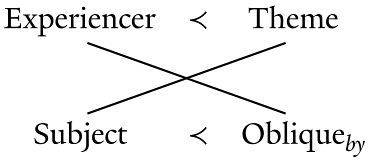


This makes it clear that in active transitive sentences, subject experiencer verbs display aligned linking behavior, and object experiencer verbs display misaligned linking behavior. In our diagrams, misalignment is visually apparent as a crossing of the lines that link each thematic role to the grammatical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

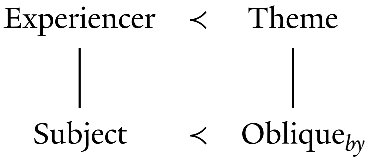
1.2 Psycholinguistic approaches to thematic role alignment

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.[4] To show how this works, consider the linking diagrams from (3), modified for passive uses of different subclasses of experiencer verbs, shown in (4).

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



b. Object experiencer passives (aligned):



As (3–4) 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

[4] Note that this move takes us beyond most formal syntactic and semantic approaches to thematic role alignment, whose primary goal is to account for the regularity of linking behavior in underlying configurations.

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).

This hypothesis makes a range of predictions for the comprehension and production of experiencer verbs given (3–4). In particular, active subject experiencer verbs should be easier to process than (i) active object experiencer verbs and (ii) passive subject experiencer verbs, and passive object experiencer verbs should be easier to process than (iii) active object experiencer verbs and (iv) passive subject experiencer verbs. If the same constraints underlie production as well, we similarly predict easier production and relatively higher frequency of aligned configurations compared to misaligned configurations (cf. Do & Kaiser 2019, 2022; Gennari & MacDonald 2009).

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., Hammerly 2020; Hammerly et al. 2022; Traxler et al. 2002; Wagers & Pendleton 2016). For ease of exposition, we refer to the specific hypothesis that configurations that align thematic roles with grammatical functions should be easier to produce and process than misaligned linking configurations as the **thematic-grammatical alignment hypothesis**. While we do not discuss all studies that have examined alignment's role in processing and production in detail in order to focus on experiencer verbs specifically, proposals regarding such alignment effects have differed in where they locate their source: comprehension, production, or factors underlying both.

1.2.1 Alignment in comprehension

Bornkessel & Schlesewsky (2006) and Bornkessel-Schlesewsky & Schlesewsky (2009) propose that alignment across many different kinds of prominence scales is important to processing. Their extended Argument Dependency Model (eADM) is an example of a model that directly locates the effects of alignment in comprehension. They propose that comprehenders predict thematic roles for arguments greedily, and prefer to link arguments encountered earlier to higher thematic roles. However, 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.

Most relevant to the thematic-grammatical alignment hypothesis, they discuss evidence from Bornkessel et al. (2002, 2003) that there is a preference to align thematic roles with morphological case marking. Their evidence comes from ERP studies of German sentences containing dative experiencer verbs like the following:

(5) ... dass der Dirigent den Sängerinnen auf-fällt.
... that the conductor.NOM the singers.DAT on-fall.3SG.PRES
'... that the conductor is striking to the singers.'
(Bornkessel & Schlesewsky 2006 (15))

Such sentences involve a misalignment between a thematic role scale (*Experiencer* ≺ *Theme*) and a morphological case marking scale (*NOM* ≺ *DAT*). Comprehenders initially prefer to assign a higher thematic role to the nominative argument than to the dative argument, but must revise this to integrate the meaning of the verb. This reanalysis can show up as an early positivity in the parietal region (Bornkessel et al. 2002, 2003). Thus, while their proposal that alignment across hierarchies is important to processing is partially in line with the thematic-grammatical alignment hypothesis, their proposal differs from it in important ways. In particular, Bornkessel & Schlesewsky (2006) explicitly reject the use of a grammatical function hierarchy in favor of a morphological case hierarchy, since there is no evidence for effects of grammatical function in the data they discuss from German, and the English data they discuss is consistent with a linear order preference (see section 1.2.3).

Nevertheless, several studies have directly investigated thematic-grammatical alignment in online comprehension, though patterns of effects have not always been consistent. 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. In a plausibilty judgment study, she investigated how quickly and accurately participants rejected implausible sentences containing agent-theme, subject experiencer, and object experiencer verbs with one animate argument and one inanimate argument. In her items, the source of the implausibility was due to the inanimate argument being linked to the agent/experiencer role, as in (6–8).

(6) Agent-theme:
 a. The signature refused the supplier. (active)
 b. The supplier was refused by the signature. (passive)

(7) Subject experiencer:
 a. The tunnel liked the youngster.
 b. The youngster was liked by the tunnel.

(8) Object experiencer:
 a. The messenger convinced the diagram.
 b. The diagram was convinced by the messenger.

Participants were slower to respond to active object experiencer verbs than the actives of the other two verb types combined, and made fewer errors in plausibility judgments for passives of object experiencer verbs compared to the other verb types. In an additional stops-making-sense experiment, she found that active object experiencer verbs led to longer RTs than active subject experiencer verbs on the first word following the verb, though there was only a marginal effect on error rates. For passives, object experiencer verbs led to fewer errors at the critical verb compared to subject experiencer verbs; and object experiencer passives led to faster responses and fewer errors than subject experiencer passives on *by*. A final experiment revealed that subject experiencer passives led to greater error rates compared to the combined set of object experiencer verbs and matched adjectival predicates at the critical predicate, but no effect was found on response time. At the *by* in passives, she found that object experiencer passives led to shorter RTs compared to subject experiencer passives, but there was no difference found on error rates. Thus, while for each word there was never evidence of effects on both RTs and error rates, overall her results found that in sentences with one animate and one inanimate argument, misaligned configurations were associated with higher RTs and lower accuracy, consistent with the thematic-grammatical alignment hypothesis. However, an earlier study with a similar design failed to yield clear evidence for the thematic-grammatical alignment hypothesis (Carrithers 1989).

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.[5] Sample items for the relevant conditions are in (9).

(9) 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)

Responses to comprehension questions for ORCs with active object experiencer verbs were similarly less accurate than for other conditions. However, there is a potential structural confound in their design (i.e., comparing ORCs to SRCs). Manouilidou & de Almeida (2013) compared active sentences with object experiencer and subject experiencer verbs, and 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, with only the greater difficulty for active object experiencer verbs consistent with the predictions of the thematic-grammatical alignment hypothesis. Nevertheless, other studies using a variety of measures including SPR, eye-tracking while reading, and MEG have generally found results consistent with the thematic-grammatical alignment hypothesis (Brennan & Pylkkänen 2010; Paolazzi 2018; Paolazzi et al. 2019, 2021).

Converging evidence for the thematic-grammatical alignment hypothesis can also be found in 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 the predictions of the thematic-grammatical alignment hypothesis, 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).

In sum, previous work on alignment in comprehension has generally found evidence consistent with the thematic-grammatical alignment hypothesis. However, comprehension results are not en-

[5] Gennari & MacDonald (2008) did not interpret their results this way, because their study was designed to cross voice (active, passive) with the animacy of the head noun of the relative clause (animate, inanimate). However, most of their items in the animate head noun conditions used object experiencer verbs, so their results can alternatively be interpreted as generally consistent with the notion that active object experiencer verbs (for them, active animate head noun condition) produce longer RTs than passive object experiencer verbs (for them, passive animate head noun condition), which is in line with the predictions of the thematic-grammatical alignment hypothesis. Indeed, Gennari & MacDonald (2009) reanalyze the RTs from Gennari & MacDonald (2008) in their study 5, with verb type explicitly considered relevant to the pattern of results.

tirely consistent across studies, and few studies on non-impaired adult native speakers have fully crossed voice and verb type, with results of these studies sometimes being internally inconsistent.

1.2.2 Alignment in production

While the eADM makes predictions for the processing of different kinds of alignment configurations, it does not directly do so for production, under the assumption that production starts at a message level (Bock & Levelt 1994) and thus does not require reanalysis in misaligned cases. An alternative explanation of difficulties associated with alignment is found in Wagers & Pendleton (2016) and Wagers et al. (2018), who propose modeling on-line effects related to (mis)alignment of distinct prominence hierarchies using a optimality-theoretic system (Prince & Smolensky 1993), which makes use of ranked constraints (see also van Bergen 2011, for a similar proposal in the context of modeling word order variation in Dutch). Markedness constraints for misaligned configurations are ranked higher than ones for aligned configurations, leading to a preference for aligned configurations when all else is equal. Wagers & Pendleton (2016) propose modeling the facilitating effect of inanimate heads selectively on ORCs (see also Baird & Koslick 1974; Caplan & Waters 1992; Ford 1983; Friederici et al. 1998; Gennari & MacDonald 2008, 2009; Gordon et al. 2001, 2004; King & Just 1991; MacWhinney & Pléh 1988; Mak et al. 2002, 2006; Staub 2010; Staub et al. 2017; Traxler et al. 2002; Warren & Gibson 2002; Waters & Caplan 1996a,b; Waters et al. 1987) as reflecting a universal preference to align a grammatical function hierarchy *Subject ≺ Object* and an animacy hierarchy *Animate ≺ Inanimate*, which in some languages rises to the level of a grammatical requirement (e.g., Chamorro: Chung 2012; Clothier-Goldschmidt 2015; see also Keenan & Comrie 1977). They implement this idea using four markedness constraints, ranked as follows ('*' indicates that a configuration is dispreferred):

(10) a. *Subject|Inanimate ≺ *Subject|Animate
(i.e., inanimate subjects are more dispreferred than animate subjects)

b. *Object|Animate ≺ *Object|Inanimate
(i.e., animate objects are more dispreferred than inanimate objects)

These constraints penalize exactly the kind of misaligned configurations we have shown in previous diagrams. Wagers & Pendleton propose that comprehenders posit representations that maximize well-formedness according to such constraints; as such, when the head of a relative clause is inanimate, comprehenders are less likely to posit a subject gap, and the increased difficulty of ORCs is attenuated (see also Wagers et al. 2018). Additional evidence for this proposal comes from He & Chen (2013), who find similar effects in Mandarin pre-nominal ORCs, and Hammerly (2020) and Hammerly et al. (2022), who find evidence from a study of Border Lakes Ojibwe (Nishnaabemwin) for effects related to alignment of personhood and grammatical function scales. A similar proposal specifically for production can be found in Kaiser et al. (2011), who found that misaligned (i.e., passive agent-theme) structures led to increased competition between possible referents in a sentence continuation task measuring pronominal reference resolution.

The crucial difference between Bornkessel & Schlesewsky (2006)'s eADM and Wagers & Pendleton (2016) and Wagers et al. (2018)'s constraint-based approach lies in the proposed source of misalignment penalties. In the eADM, effects of misalignment arise during incremental processing. In contrast, the constraint-based approach identifies the problem as inherent to the grammatical representation of misaligned configurations. As such, the constraint-based approach more straightforwardly predicts production effects of alignment that mirror ones found in comprehension, under the assumption that the same grammatical representations underlie both modalities (e.g., Momma & Phillips 2018). In contrast, comprehension-oriented models such as the eADM do not necessarily predict any effects of misalignment during production.

Effects of thematic-grammatical alignment have, in fact, been widely observed in sentence production, which suggests that the representation of alignment plays a role in language use that underlies both production and processing. For example, Ferreira (1994) showed that speakers were relatively more likely to produce 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 weakened when both arguments were animate. She also found that time to speech onset for active sentences with object experiencer verbs displayed a greater penalty when the arguments differed in animacy compared to when they did not, with a smaller penalty for this difference with agent-theme and subject experiencer verbs. Altmann & Kemper (2006) (see also Altmann et al. 2004) found similar results in their production experiments, though unlike Ferreira (1994)'s results, they found no difference in speech onset latency between passives and actives for object experiencer verbs.

Do & Kaiser (2019, 2022) examined how thematic-grammatical 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. They found that 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. Such effects were not found when participants merely examined the images but did not produce descriptions, supporting a view that their production results were related specifically to linguistic encoding. Frequency of verb use in active or passive voice was not a significant predictor of their effects. Thus, while Ferreira (1994) and Altmann & Kemper (2006) show consistent effects of alignment on choice of structure (active or passive), Do & Kaiser (2019, 2022)'s results show that alignment influences on-line processes in production planning, even when structure choice is forced.

1.2.3 An alternative to grammatical function alignment: Experiencer-first

The preceding review suggests that there is a good reason to suppose that thematic-grammatical alignment affects both production and comprehension processes. However, an alternative interpretation of the results we have heretofore discussed is possible in terms of an experiencer-first preference. In English simple active and passive sentences, which most of the studies we discussed used,[6] linear order and grammatical function are confounded: subjects go first and objects go second. Thus, alignment of thematic roles with a linear order hierarchy *First* $\prec$ *Second* would predict the same results as the thematic-grammatical alignment hypothesis. Such a preference could be due to cognitive constraints (e.g., Gennari & MacDonald 2009), as thematic roles that tend to refer to relatively more salient arguments (e.g., experiencers compared to themes) may lead to easier retrieval during production and easier processing in earlier positions.

Additional work has indeed found evidence consistent with an experiencer-first preference. Ferreira (2003) compared the comprehension of simple actives and passives to comprehension of subject and object clefts. She found comprehenders were less accurate in linking thematic roles to argu-

[6] Gennari & MacDonald (2008, 2009) used SRCs and ORCs, but they interpreted their results as reflecting an experiencer-first preference given the specifics of their experimental design.

ments in passive sentences than in active sentences. This was compared to performance in subject clefts (*It was the S that Ved the O*), which are less frequent than actives but display canonical linking properties, and object clefts (*It was the O that the S Ved*), which are also less frequent but place the theme argument before the agent argument (with agent-theme verbs). Frequency did not affect ease of comprehension and reading times, while linear order alignment did, with OS orders leading to fewer accurate responses.

While Ferreira (2003)'s results have an potential structural confound – namely, the use of clefts, which may carry additional difficulties – evidence for an experiencer-first preference from languages other than English is more compelling. 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)[7]; Italian: Dröge et al. 2014 (ERP); Spanish: Gattei et al. 2015a[8] (SPR, acceptability judgment), Gattei et al. 2015b (ERP), Gattei et al. 2017, 2022[9] (eye-tracking while reading); L1 Spanish: Mateu 2022 (picture-matching)).

Thus, a credible alternative to the thematic-grammatical alignment hypothesis could assume an experiencer-first preference, with grammatical functions being relevant only to the extent that English confounds them with linear order. Nevertheless, thematic-grammatical alignment preferences may yet be active even in languages that allow simple OS clauses. 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, while there may be an experiencer-first preference when a given sentence is comprehended, there may be both experiencer-first and thematic-grammatical alignment preferences that are revealed during normal production, since speakers can choose which grammatical function to assign to a particular argument, as well as which argument comes first.

1.2.4 Frequency as the link between production and comprehension

One way of interpreting a possible experiencer-first comprehension preference is as reflecting cognitive constraints that most directly affect production (Gennari & MacDonald 2009; MacDonald

[7] In Kretzschmar et al. (2012)'s study, the effect was contingent on ambiguous case marking and limited to sentences where the initial argument in the relative clause containing the verb was a bare plural.

[8] Gattei et al. (2015a,b) and Gattei et al. (2017) consistently refer to Spanish verbs like *gustar* 'be pleasing' as object experiencer verbs. However, according to certain syntactic analyses (Cuervo 2003, 2010, 2020), these may in fact be subject experiencer verbs that mark their subjects with dative case. We note that if these verbs should be analyzed as having dative subjects rather than dative objects, the interpretation of their results may be more consistent with there being a thematic-grammatical alignment preference along with an experiencer-first preference. This is because if the verbs they used are actually dative subject experiencer verbs, then the OVS label they applied to sentences where the dative experiencer preceded the nominative theme are actually in an SVO order, and are still aligned according to an analysis like Belletti & Rizzi (1988)'s.

[9] While Gattei et al. (2022) characterize their verb types as accusative, agent-experiencer verbs and dative, theme-experiencer (i.e., object experiencer) verbs, the former might be plausibly considered to be eventive agent-theme uses of object experiencer verbs. Though we are aware of no direct claims to this effect for Spanish, it has been widely noted in the formal syntactic and semantic literature on experiencer verbs that the special linking properties of object experiencer verbs (and possibly their invocation of an experiencer argument) is restricted to stative uses (cf. Belletti & Rizzi 1988; Hirsch 2018; Landau 2002, 2010; Pesetsky 1995). However, as long as the accusative condition invokes an agent, this difference would not substantially affect the interpretation of their results, since experiencers are ranked below agents in all thematic hierarchies that distinguish them to our knowledge.

2013). Gennari & MacDonald (2009) and MacDonald (2013) analyze the experiencer-first preference this way in their Production-Distribution-Comprehension (PDC) framework. According to PDC, cognitive constraints affect production preferences, which lead to distributional properties in the input comprehenders receive. Higher frequency configurations are processed more easily due to comprehenders' greater experience with them. A large body of work (e.g., Dell & Chang 2014; Gennari & MacDonald 2009; Gennari et al. 2012; Hsiao & MacDonald 2016; Keshev & Meltzer-Asscher 2021; Levy 2008a,b; MacDonald 1999, 2013; MacDonald et al. 1994; MacDonald & Thornton 2009; Trueswell 1996; Wells et al. 2009; Wonnacott et al. 2008 , a.o.) shows that frequency information plays an important role in comprehension and (error-based) learning.

Most germane to the topic of experiencer verbs is a paper by Gennari & MacDonald (2009) that lays out PDC and its relevance for different comprehension behavior of agent-theme and object experiencer verbs. In two production studies, Gennari & MacDonald (2009) elicited RCs with object experiencer verbs and agent-theme verbs. They found that more active ORCs were produced for agent-theme verbs, while for object experiencer verbs a preference for more passives was modulated by the animacy of the RC's head noun. The production preferences observed in their experiments were matched by counts of the same structures in the Wall Street Journal parsed corpus and the British National Corpus. To establish a link to comprehension, they regressed individual verbs' passivization rates on residual RTs from experiment 2 of Gennari & MacDonald (2008), which examined active ORCs of animate-subject inanimate-object agent-theme verbs and inanimate-subject animate-object object experiencer verbs. They found that higher frequency structures were easier to process, with similar results in an additional SPR study that manipulated animacy.

However, while Gennari & MacDonald (2009) examined actives and passives of agent-theme and object experiencer verbs in their production and corpus studies, they did not examine passives in their SPR studies, nor did they examine subject experiencer verbs in either. Thus, their results do not directly provide evidence bearing on the question of whether active and passive linking behavior is reversed in comprehension (which is relevant not only to the thematic-grammatical alignment hypothesis). Our study addresses these gaps and concerns by examining subject and object experiencer verbs in active and passive structures.

Secondarily, while Gennari & MacDonald (2009) and other prior work has explored the role of production frequency in predicting reading time data, we are not aware of work on this topic reporting how models fit using frequency alone compare to models fit using categorical grammatical predictors or to models fit using both. This is potentially important if we wish to compare PDC to other frameworks. PDC proposes a direct causal link between frequency and comprehension. However, it is also possible that frequency is related to comprehension only indirectly. This would be consistent with a model where grammatical constraints influence production, which influences distributional frequencies, while those same grammatical constraints directly influence comprehension, with frequency having a more limited effect. Under this approach, comprehension *could* be predicted from frequency alone, due to the indirect link between frequency and comprehension. However, we would expect that grammatical constraints would be better predictors of comprehension behavior than frequency. Thus, a secondary aim of our analyses is a comparison of how models fit using frequency alone compare to models using grammatical predictors relevant to alignment, in order to evaluate the relationship between frequency, grammatical constraints, and comprehension.

1.3 The present study

As discussed 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 (such as animacy, linear order, voice, and verb type). The main goal of our study is to build on this previous work to provide a direct test of the thematic-grammatical alignment hypothesis. Our study aims to fill the relevant gap in the prior literature with stimuli like the following.

(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)

The predictions of the thematic-grammatical alignment hypothesis are straightforward: active subject experiencer configurations should be easier to process than passive subject experiencer configurations, but the reverse should be true for object experiencer configurations. Likewise, active subject experiencer configurations should be easier to process than active object experiencer configurations, but the reverse should be true for passives.

The examples in (11) use ORCs. This means that experiencer-first accounts make exactly the opposite prediction compared to the thematic-grammatical alignment hypothesis: If the first noun (i.e., the head of the relative clause) is preferentially assigned an experiencer role, then active subject experiencer and passive object experiencer configurations should be more difficult. To further illustrate this contrast, the relevant configurations under each linking hypothesis are presented in figure 1.

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 makes it easier to measure the difficulty of this process using word-by-word processing measurements.

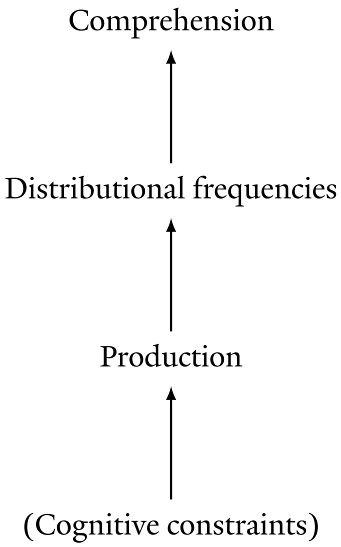
In addition, the fact that the head of the ORC was the main clause object meant that the head noun had the same grammatical function it had in the embedded clause in the main clause (in actives), or at least a more comparable one (in passives, because objects are closer to obliques on the grammatical 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.

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 the opposite prediction of the thematic-grammatical alignment hypothesis.

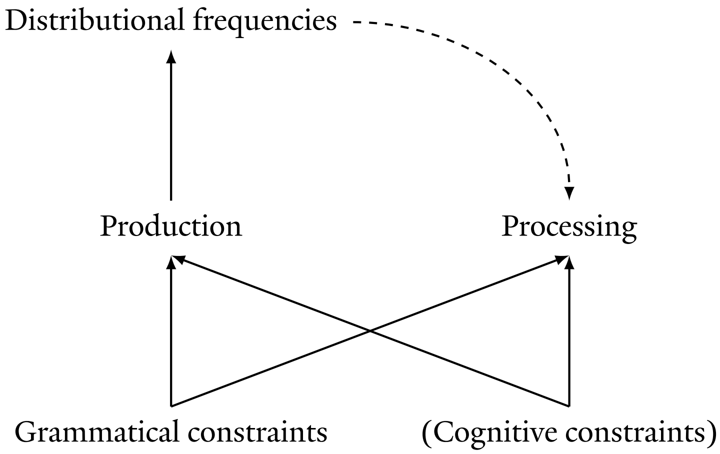
A secondary goal of the present study is to probe how linking might impact comprehension. One possibility is that linking only influences comprehension indirectly. This view is suggested by the PDC framework, which highlights the role of statistical generalization over language experience in tuning probabilistic constraints or cues that drive language comprehension (a perspective it broadly shares with constraint-based approaches to language processing (MacDonald & Seidenberg 2006)). In the PDC framework, linking principles influence how producers formulate their utterances, which in turn affects distributional frequencies. Learning to comprehend language involves tuning probabilistic cues to comprehension on the basis of these distributions. In this way, the comprehender's behavior is influenced indirectly by production pressures via frequency distributions, which in turn affect processing/comprehension via statistical learning processes. The causal relations on this view are summarized in figure 1, which highlights the key role that distributional frequencies play in determining the comprehender's preferences (for more in-depth discussion, see MacDonald 2013).

However, it could also be that linking principles more directly apply in production and processing based on how these latter processes must interact with different parts of the grammar (cf. Hammerly et al. 2022; Momma & Phillips 2018; Wagers & Pendleton 2016; Wagers et al. 2018). This view is also broadly consistent with constraint-based approaches to comprehension, though it emphasizes the possibility that grammatical constraints – such as the relative markedness of different linking configurations – may directly be activated to guide both production and comprehension processes. The different causal relations this view supposes are also summarized in figure 1. We refer to this as the "Grammar Underlies Production and Processing (plus frequency)" (GUPPy) framework (see also the "Maximize Incremental Well-formedness" proposal in Wagers et al. 2018).

(12) a. PDC (based on Gennari & MacDonald 2009; MacDonald 2013):



b. GUPPy:



The PDC framework posits that experience (i.e., distributional frequencies) directly influences comprehension; grammatical constraints in PDC influence comprehension only indirectly. In contrast, GUPPy models grammatical constraints (in interaction with cognitive constraints) as directly influencing both production and processing. While GUPPy builds in PDC (one can trace the same causal path that PDC describes), it models grammatical constraints as having a direct role in production and processing. GUPPy is thus able to account for frequency effects and the importance of probabilistic cue weighting during comprehension processes, but is consistent with grammatical constraints having a more direct link to processing.

The PDC and GUPPy frameworks helpfully summarize two existing views on the role that linking pressures can play in language comprehension. Importantly for our present purposes, these frameworks imply different causal models of how grammatical pressures and frequency can predict behavioral processing measures (see figure 1). We capitalize on this perspective in our studies below, and use Bayesian modeling to attempt to directly compare statistical models with grammatical predictors to models with frequency predictors. The PDC framework would most straightforwardly predict that the frequency-based model should better predict processing measures than the grammatical predictors model, since it is frequency alone that is directly linked to processing. In contrast, the GUPPy framework would predict that grammatical predictors should be more tightly linked to processing than frequency.

1.4 Summary

In sections 1.1–1.2, we reviewed Belletti & Rizzi (1988)'s approach to the formal syntax and semantics of experiencer verbs, as well as psycholinguistic evidence that is compatible with an experiencer-first preference as well as the thematic-grammatical alignment hypothesis. Figure 1 presents a summary of the alignment configurations we discussed, as well as the different psycholinguistic explanations for the effects that have been found and the predictions each makes.

2. Experiment 1

Experiment 1 aimed to test the thematic-grammatical alignment hypothesis, according to which thematic roles' alignment with grammatical functions should facilitate processing and comprehension. According to GUPPy, this is because comprehenders should be able to more easily recover the correct underlying representations when these configurations are transparently reflected in the form of the input than when they are not.

To be more specific, the thematic-grammatical alignment hypothesis 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, we might predict effects of voice nested with Verb type, as well as effects of Verb type nested within voice.

2.1 Materials

2.1.1 Stimuli

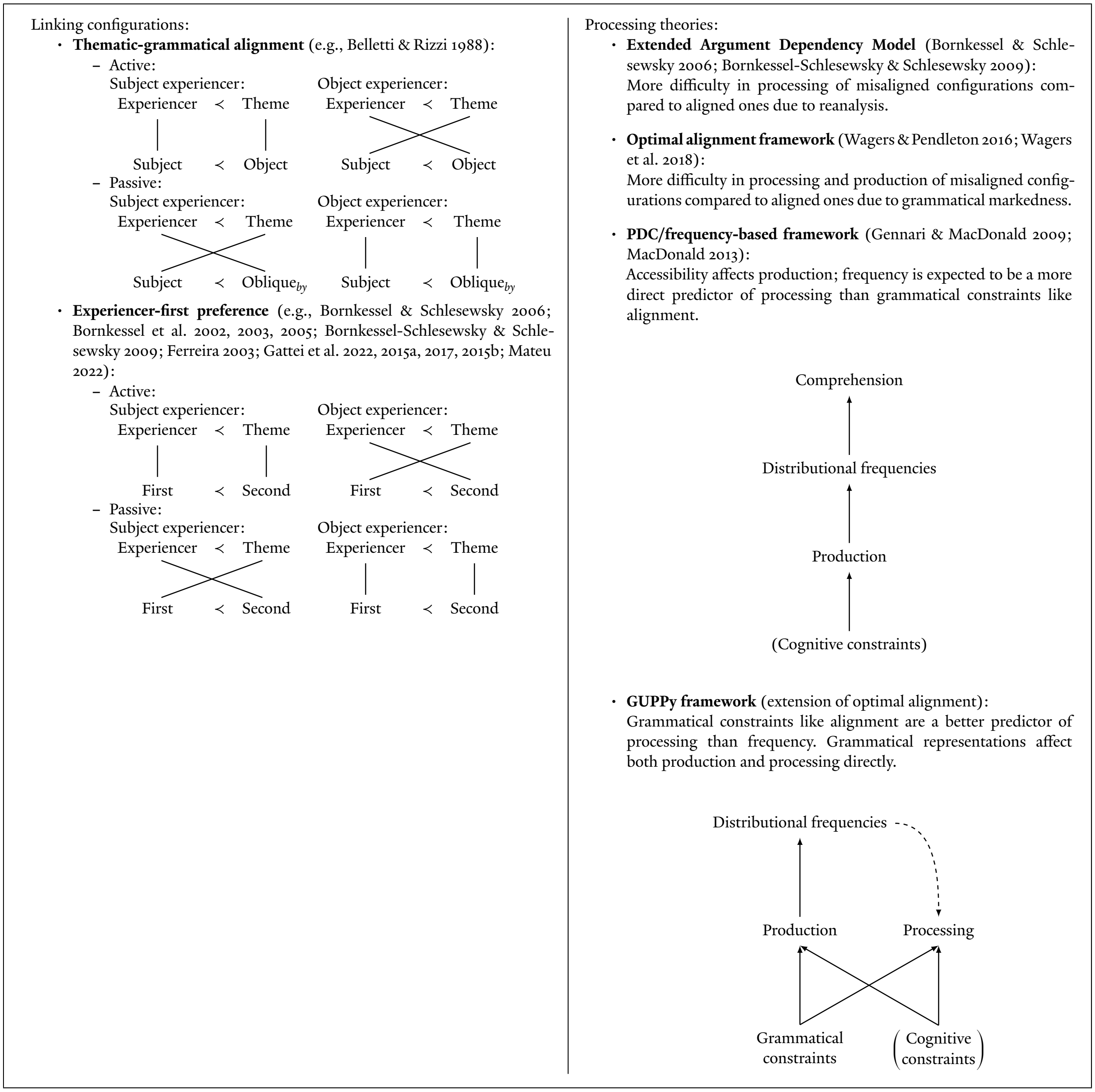
We constructed 32 items like those in (11) in a within-subjects 2 × 2 design, crossing Verb type (subject experiencer, object experiencer) and Sentence voice (active, passive). We used 32 verbs (16 subject experiencer, 16 object experiencer), based on Levin (1993)'s categorization[10] (a full list of items can be found in Appendix A).

(13) 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 $t(29.96) = 0.03, p = 0.98$).

[10] 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.'



Linking configurations:

- **Thematic-grammatical alignment** (e.g., Belletti & Rizzi 1988):
  - Active:
  - Passive:

- **Experiencer-first preference** (e.g., Bornkessel & Schlesewsky 2006; Bornkessel et al. 2002, 2003, 2005; Bornkessel-Schlesewsky & Schlesewsky 2009; Ferreira 2003; Gattei et al. 2022, 2015a, 2017, 2015b; Mateu 2022):
  - Active:
  - Passive:

Processing theories:

- **Extended Argument Dependency Model** (Bornkessel & Schlesewsky 2006; Bornkessel-Schlesewsky & Schlesewsky 2009):
  More difficulty in processing of misaligned configurations compared to aligned ones due to reanalysis.

- **Optimal alignment framework** (Wagers & Pendleton 2016; Wagers et al. 2018):
  More difficulty in processing and production of misaligned configurations compared to aligned ones due to grammatical markedness.

- **PDC/frequency-based framework** (Gennari & MacDonald 2009; MacDonald 2013):
  Accessibility affects production; frequency is expected to be a more direct predictor of processing than grammatical constraints like alignment.

- **GUPPy framework** (extension of optimal alignment):
  Grammatical constraints like alignment are a better predictor of processing than frequency. Grammatical representations affect both production and processing directly.

Figure 1: Summary of linking configurations and predictions of processing theories

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 & D: subjexp = 7.81, objexp = 8.31; Welch's $t(29.83) = -0.95, p = 0.35$; mean groups B & 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 (in a region-based analysis that can found in Appendix E) and to contain a spillover word for actives. 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 (11)):

(14) 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)

(15) 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. 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.

Items were split across 16 lists using a Latin Square design such that no participant saw each item in more than one condition. Lists were arranged such that 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 Corpus study

Since a secondary aim of our study was to compare effects of frequency to effects of alignment, we conducted a corpus study, following Gennari & MacDonald (2009). The specific question we addressed was 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 give the best possible chance to approaches where frequency has the most direct effect on processing. 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).[11]

[11] 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 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

Next we considered the relative frequency of active and passive structures across each verb. Figures 2 and 3 show the percentages (on the *y*-axis) and counts (above each bar) including and excluding probable adjectives, respectively.

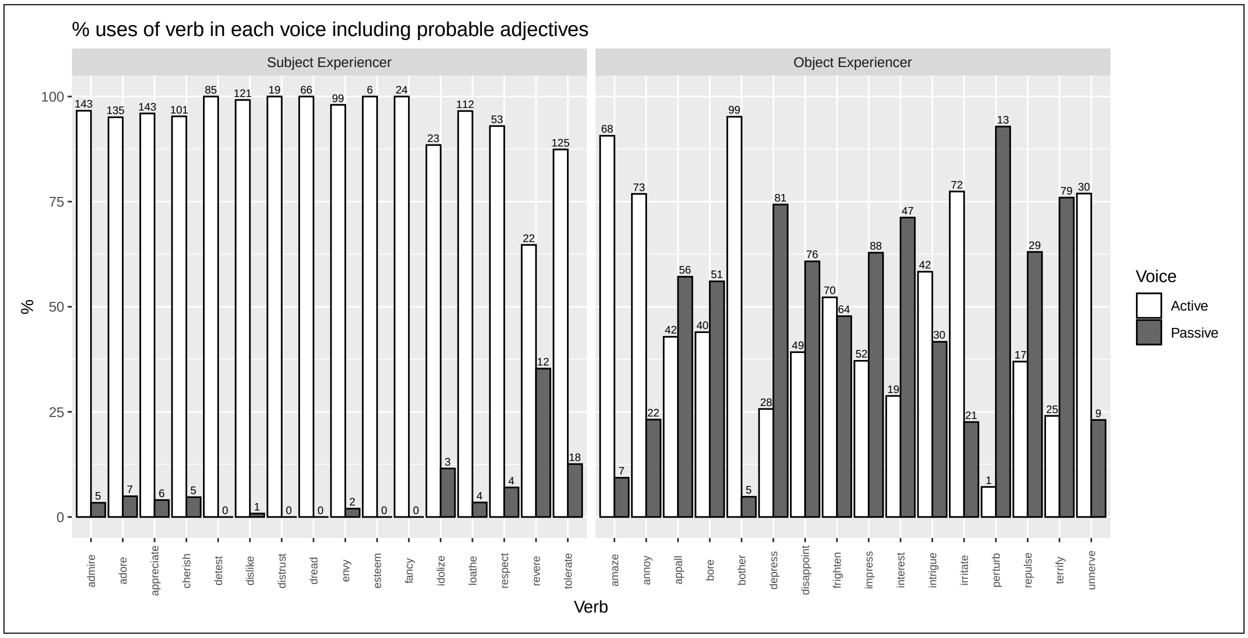


Figure 2: Frequency of voice by verb including probable adjectives

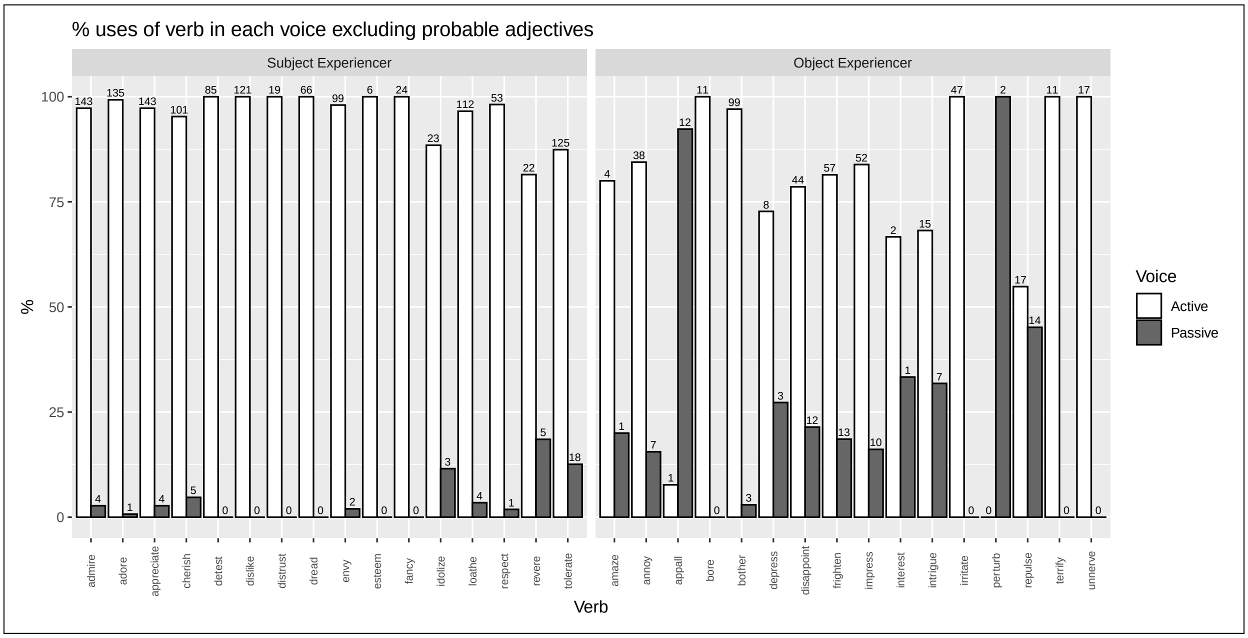


Figure 3: Frequency of voice by verb excluding probable adjectives

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.

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 (16),

(16) $$\log\left(\frac{n_{\text{Verb,Voice}_i} + 0.5}{n_{\text{Verb,Voice}_j} + 0.5}\right)$$

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 observed voice compared to the unobserved voice for a particular trial. 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.

2.1.3 *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 presented 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.

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.[12] We also considered question accuracy. All RTs and question responses from eligible participants were analyzed.[13]

[12] 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 interpretation of the results.

[13] In addition to the analyses based on word log RTs, we also conducted region-based analyses, which are described

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` library (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. 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 (including frequency), interactions, and σ, we used priors of $\mathcal{N}(0, 1)$. For the correlation matrix prior, we used $\mathrm{LKJ}(2)$. For RTs, we fit multilevel MCMC models on the critical verb and the spillover word.

To quantify the strength of evidence for the effects in our study, we summarize the posterior distributions over key model parameters using p_{MCMC} values (cf. Bidgood et al. 2020). A p_{MCMC} value is the proportion of samples on the Markov chain that lie either above (for negative effects) or below (for positive effects) 0. They can be interpreted straightforwardly as the (one-tailed) probability that the effect in question does not differ from 0 (i.e., a lower p_{MCMC} value indicates greater evidence for an effect). While it is possible to interpret Bayesian models in frequentist terms—i.e., a $p_{MCMC} < 0.05$ is "statistically significant"—one can simply take the p_{MCMC} at face value and interpret the result according to their best judgment. We did not calculate p_{MCMC} values for the intercept for any model.

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), $\mu = 67.8\%$, $\sigma = 5.63\%$).

First, 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 consistent with the predictions of the thematic-grammatical alignment hypothesis. Then, we consider models based solely on frequency. We additionally fit models that included both frequency and our categorical predictors; however, the pattern of effects was largely similar to the models fit using only the categorical predictors. As such, for reasons of space, we report which effects these models found evidence for in the summary (section 2.4.3) in brief, with a more detailed presentation in Appendix B. We then report the results of model comparisons conducted to determine which models best explained the results.

2.4.1 *Voice and Verb type models*

Figure 4 displays word-by-word log RTs, with the x-axis aligned at the critical verb.[14] In all plots, bars represent standard errors.

in Appendix E. However, due to the fact that there were different numbers of words in the same region across active and passive sentences, interpreting these results is less straightforward. Nevertheless, the results of the region-based analyses are consistent with those of the word-based analyses.

[14] 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.

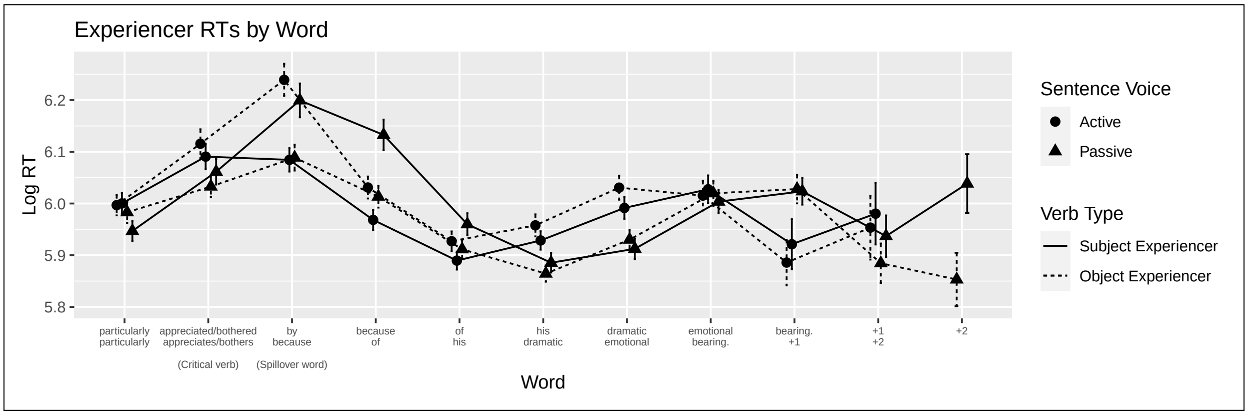


Figure 4: Log RTs by word, aligned at critical verb (Experiment 1)

Question response accuracy is presented in figure 5.[15]

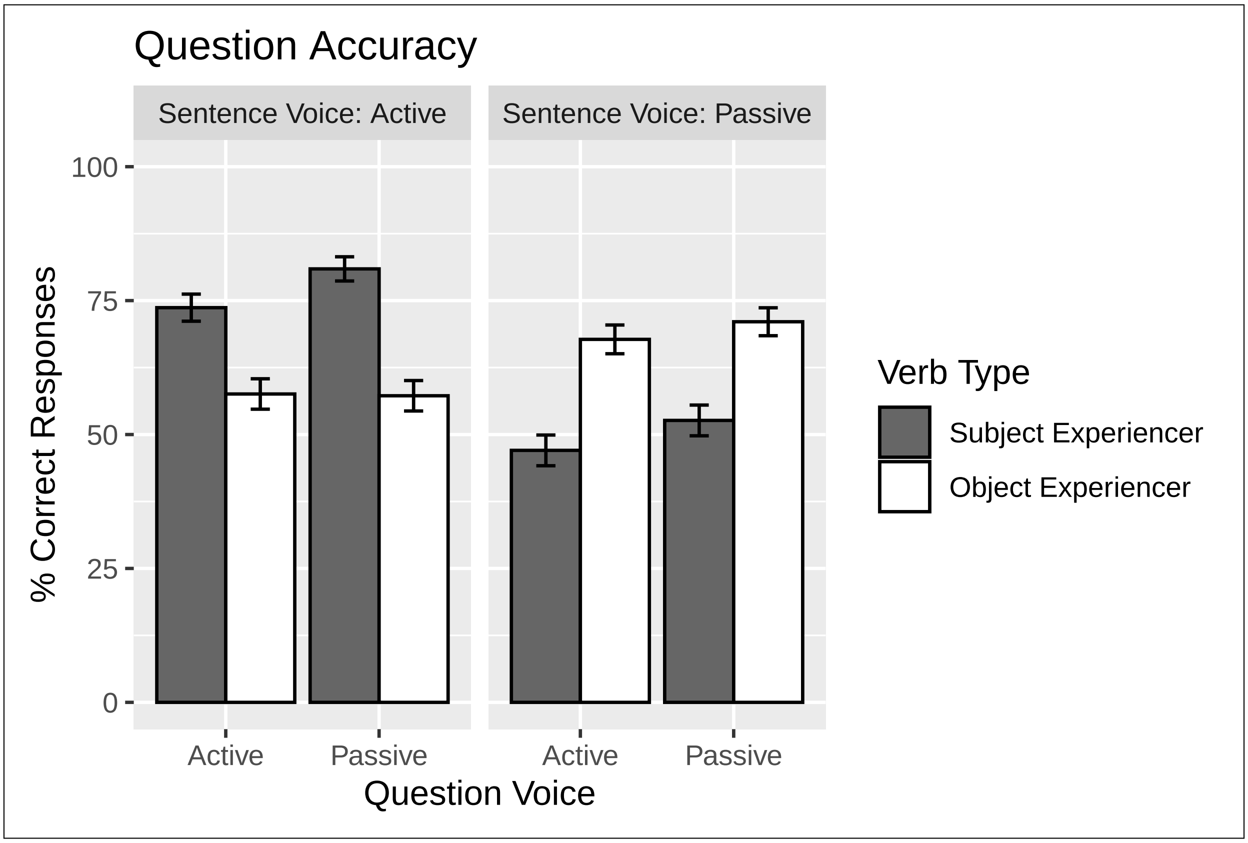


Figure 5: Question accuracy by Verb type, Sentence voice, and Question voice (Experiment 1)

[15] 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.

Critical verb At the critical verb, there was evidence for a main effect of Sentence voice (95% CI: $[-0.10, -0.01]; p_{MCMC} < 0.00$). This reflected longer overall reading times in the active condition than in the passive condition. There was no evidence for an effect of Verb type or an interaction between Verb type and Sentence voice at the critical verb, as shown in table 2 (all 95% CIs span 0).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.08	0.05	5.98	6.17	–
Verb type	−0.00	0.02	−0.04	0.04	0.55
Sentence voice	−0.06	0.02	−0.10	−0.01	< 0.00
Verb type × Sentence voice	−0.05	0.04	−0.13	0.03	0.10

Table 2: 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.

Spillover word At the spillover word, an interesting pattern emerged, as shown in table 3. There was no evidence for any main effect of Verb type or Sentence voice. Nevertheless, there was clear evidence for an interaction between these two factors (95% CI: $[-0.39, -0.14]; p_{MCMC} = 0$). We resolved this interaction by both Verb type and Sentence voice.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.15	0.05	6.05	6.26	–
Verb type	0.02	0.03	−0.03	0.07	0.27
Sentence voice	−0.02	0.03	−0.08	0.04	0.19
Verb type × Sentence voice	−0.27	0.06	−0.39	−0.14	0.00

Table 3: Crossed model for spillover word without frequency (Experiment 1)

The model with Verb type nested in Sentence voice (table 4) showed that Verb type had effects in both active and passive sentences, but in opposite directions.[16] 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 that did subject experiencer verbs.

[16] 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	p_{MCMC}
Verb type (active)	0.15	0.04	0.08	0.23	0.00
Verb type (passive)	−0.11	0.04	−0.19	−0.03	< 0.00

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

The model with Sentence voice nested within Verb type (table 5) showed that Sentence voice had effects in sentences with subject experiencer verbs and in sentences with object experiencer verbs, but in opposite directions. This reflected that 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.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Sentence voice (subjexp)	0.11	0.05	0.02	0.20	< 0.00
Sentence voice (objexp)	−0.15	0.04	−0.22	−0.08	< 0.00

Table 5: Sentence voice nested in Verb type model for spillover word without frequency (Experiment 1)

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. As table 6 shows, while there was evidence for a main effect of Question voice (95% CI: [−0.01, 0.42]; $p_{MCMC} = 0.03$), which reflects more accurate answers to passive questions than to actives overall, there was no evidence of any interaction between Question voice and the other factors. There was also evidence of a main effect of Sentence voice (95% CI: [−0.62, −0.18]; $p_{MCMC} < 0.00$), reflecting overall fewer accurate responses to questions about passive voice sentences. Finally, there was evidence for an interaction between Verb type and Sentence voice (95% CI: [1.49, 2.31]; $p_{MCMC} = 0$).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	0.66	0.10	0.47	0.86	–
Verb type	−0.04	0.13	−0.31	0.22	0.38
Sentence voice	−0.40	0.11	−0.62	−0.18	< 0.00
Question voice	0.21	0.11	−0.01	0.42	0.03
Verb type × Sentence voice	1.90	0.21	1.49	2.31	0.00
Verb type × Question voice	−0.24	0.24	−0.70	0.24	0.16
Sentence voice × Question voice	0.01	0.23	−0.45	0.48	0.48
Verb type × Sentence voice × Question voice	0.35	0.38	−0.39	1.11	0.18

Table 6: 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	p_{MCMC}
Verb type (active)	−1.02	0.17	−1.36	−0.67	0.00
Verb type (passive)	0.93	0.16	0.62	1.26	0.00
Verb type (active) × Question voice	−0.42	0.30	−1.02	0.17	0.08
Verb type (passive) × Question voice	−0.04	0.29	−0.61	0.54	0.45

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

There was evidence for effects of 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]; p_{MCMC} = 0); 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]; p_{MCMC} = 0). There was no evidence for an interaction of these effects with Question voice.

Considering the model that nested Sentence voice within Verb type next (table 8), we found evidence for effects of voice 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]; p_{MCMC} = 0), 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]; p_{MCMC} = 0). There was no evidence for an interaction between either of these effects and Question voice.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Sentence voice (subjexp)	−1.37	0.16	−1.68	−1.05	0.00
Sentence voice (objexp)	0.59	0.15	0.29	0.88	0.00
Sentence voice (subjexp) × Question voice	−0.18	0.27	−0.71	0.36	0.25
Sentence voice (objexp) × Question voice	0.22	0.34	−0.45	0.89	0.26

Table 8: Sentence voice nested in Verb type model for question accuracy without frequency (Experiment 1)

2.4.2 *Voice frequency models*

In addition to the models using only categorical predictors (Sentence voice, Verb type, and Question voice), we 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 (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 2.1.2 and defined as in (16). 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.

Plots showing the regression of means of our dependent variables on frequency are in figures 6 and 7. 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.

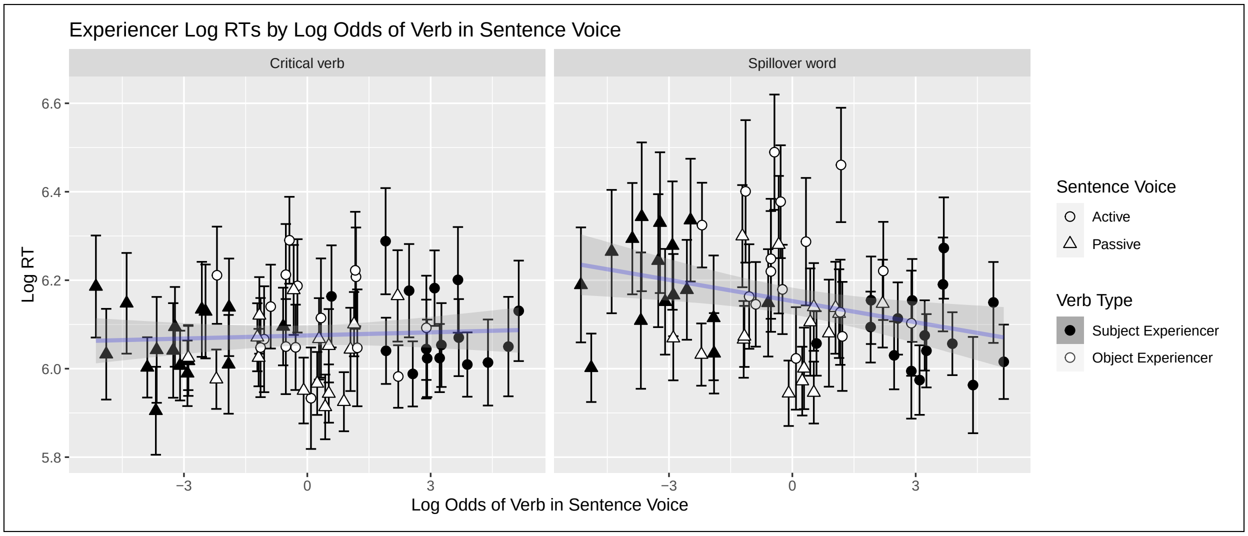


Figure 6: Regression of critical verb and spillover word mean RTs on frequency (Experiment 1)

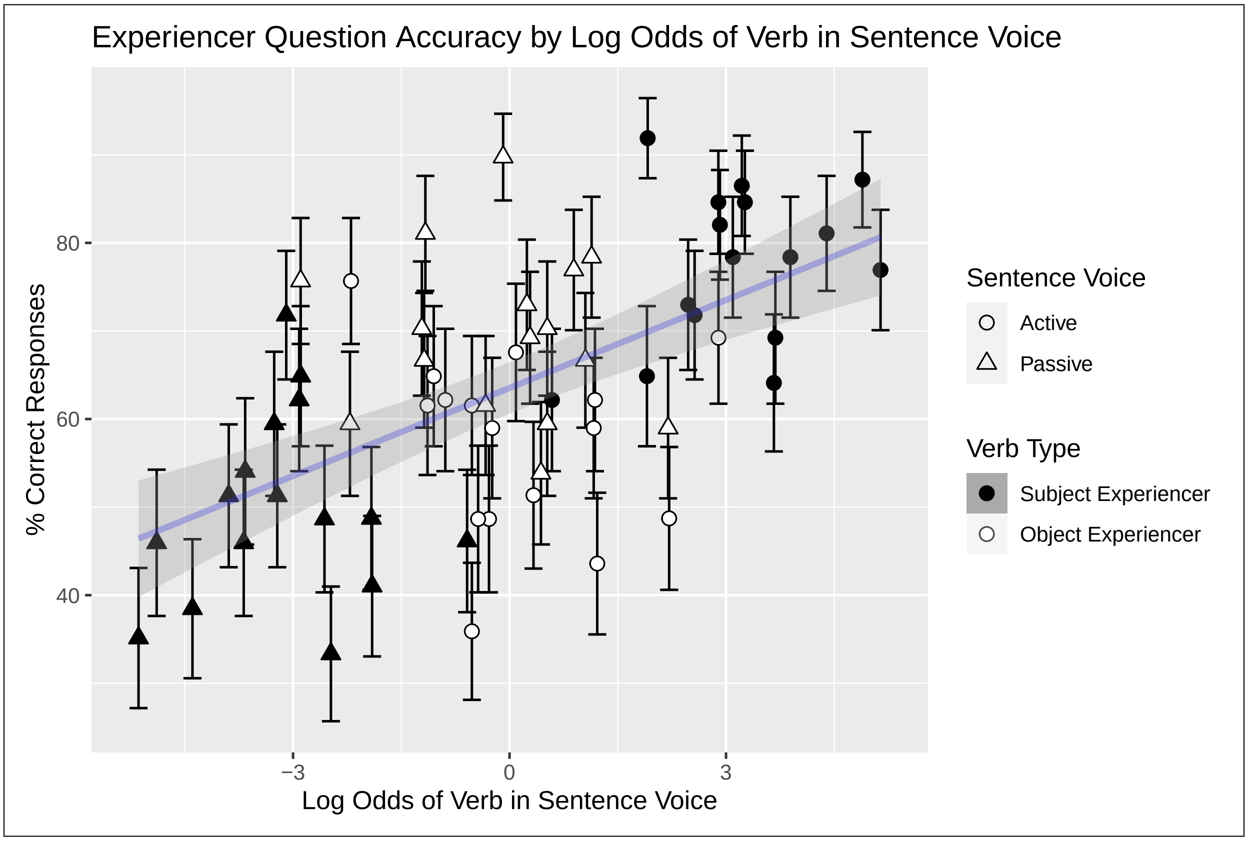


Figure 7: Regression of question accuracy on Sentence voice log odds (Experiment 1)

Critical verb There was no evidence for an effect of frequency on reading times at the critical verb (95% CI: [−0.00, 0.01]; $p_{MCMC} = 0.23$).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.07	0.05	5.98	6.17	–
Sentence voice log odds	0.00	0.00	−0.00	0.01	0.23

Table 9: Frequency model for critical verb (Experiment 1)

Spillover word At the spillover word, there was evidence for an effect of frequency (95% CI: [−0.03, −0.00]; $p_{MCMC} < 0.00$). This reflected that 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	p_{MCMC}
Intercept	6.15	0.05	6.05	6.25	–
Sentence voice log odds	−0.02	0.01	−0.03	−0.00	< 0.00

Table 10: 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 evidence for a main effect of Question voice frequency only in the Sentence voice frequency × Question voice frequency model, and no effects in any other models. This main effect went in the opposite of the expected direction, such that participants were *more* likely to answer questions accurately when the verb was *less* likely to occur in the voice in which it was presented in the question. For this reason, we leave discussion of models using Question voice frequency out of the main text. These models are presented in Appendix D.

For the question accuracy model using only Sentence voice frequency as a predictor, there was evidence for an effect of frequency (95% CI: [0.12, 0.21]; $p_{MCMC} = 0$). 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	p_{MCMC}
Intercept	0.61	0.09	0.44	0.79	–
Sentence voice log odds	0.17	0.02	0.12	0.21	0.00

Table 11: Sentence voice frequency model for question accuracy (Experiment 1)

2.4.3 *Interim summary*

Table 12 summarizes the results of the models using the different groups of predictors.

As is clear from table 12, 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 evidence for an effect of Sentence voice for subject experiencer verbs, while the model without frequency did find evidence for this effect. In contrast, while the models using only frequency as a predictor found evidence of 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 grammatical predictors find 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 2.4.4.

2.4.4 *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

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
		Question voice	active	<	passive
		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 12: Summary of model results (Experiment 1)

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.

Our comparisons used two criteria: (i) Leave-One-Out (LOO) cross-validation, and (ii) the Watanabe-Akaike Information Criterion (WAIC) (Watanabe 2010). Both types of comparisons produced similar results, so we focus on the LOO comparisons here, and present the WAIC comparison results in Appendix C. LOO cross-validation is a special case of K-fold cross-validation, where the number of folds is equal to S, the sample size. LOO thus compares models using S-fold cross-validation on the basis of the average expected log pointwise predicted density, defined as

$$(17) \qquad \widehat{\text{elpd}}\,(D, \mathcal{M}) = \frac{1}{S}\sum_{i} \log p(y_i|D_{\setminus i}, \mathcal{M})$$

where S is the number of samples, y_i is the i-th sample in the dataset D, $D_{\setminus i}$ is the dataset less the i-th sample, and $\mathcal{M}$ is the model (Nicenboim et al. 2022 (16.6)). Essentially, this means that for each sample in the dataset, the model is fit without using the information from that sample, and then the model is used to predict the log probability of that sample. These log probabilities are then averaged to give the final estimate of the expected log pointwise predicted density. We report the elpd_diff values from `brms`'s `loo` function. The elpd_diff compares the difference across models in $\widehat{\text{elpd}}$ scores relative to the model with the highest $\widehat{\text{elpd}}$ score; negative differences indicate a lower predictive value of the model compared to the one with the highest predictive value (which by definition has an elpd_diff = 0). Though the absolute LOO score for a model is uninformative because it is tied to a particular dataset, relatively more negative elpd_diff scores for LOO indicate lower predictive utility when comparing models fit to the same data.

Sivula et al. (2020) offer the following rules of thumb for interpreting elpd_diff scores: if (the absolute value of) the difference is less than 4, then the models have very similar predictive performance regardless of the standard error of the difference. Otherwise, if the absolute value of the elpd_diff is > 4, there are > 100 observations (our dataset has 2,432 observations total for each measure), and the model is not badly misspecified, then the elpd_diff and standard error of the difference can be relied upon to accurately reflect differences in the predictive values of the models, and one can use the standard error to determine how large the difference in predictive utility between the two models is.

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 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).

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.

For each model, we also report the number of problematic observations (out of 2,432 total) for the comparison, as *n* prob. obs. These are points that were identified as having a Pareto $\hat{k} > 0.7$. Such points make comparison estimates less reliable. Most of our LOO comparisons found no problematic observations.

Critical verb

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	0
SV × VT + SFq	−2.6	0.8	0
SFq	−2.2	5.0	0

Table 13: Critical verb RTs predictive crossed model LOO comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + SFq	−2.6	0.9	1
SFq	−3.2	4.9	0

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

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	2
VT + SV ∈ SE + SV ∈ OE + SFq	−3.0	0.9	0
SFq	−5.0	5.6	0

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

Predictive comparisons In general, across the predictive comparisons in the Critical verb, we found no evidence of different predictive utility: though numerically the model that did not include frequency had the highest predictive utility, the others did not differ substantially from it.

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	0
SV × VT + *SFq*	−0.8	0.7	0
SV × *VT* + SFq	0.0	0.0	0
SFq	−3.8	3.4	0
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + *SFq*	−0.8	0.7	0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	0
SFq	−3.4	3.4	0
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	2
VT + SV ∈ SE + SV ∈ OE + *SFq*	−1.7	0.8	0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	0
SFq	−3.5	3.4	0

Table 16: Critical verb RTs additive model LOO comparisons (Experiment 1)

Additive comparisons We also found no reliable differences in predictive utility for the additive comparisons: as before, adding frequency to the models without it produced no clear improvement. However, in addition, adding the categorical predictors to the frequency only model did not produce any improvement either.

Spillover word

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	0
SV × VT + SFq	−2.2	1.3	0
SFq	−17.7	8.6	0

Table 17: Spillover word RTs predictive crossed model LOO comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + SFq	−1.6	1.6	0
SFq	−13.4	8.3	0

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

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE + SFq	−0.3	1.4	0
SFq	−16.7	8.0	0

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

Predictive comparisons Unlike the comparisons for the critical verb, we found some differences in predictive utility in the predictive comparisons for the spillover word. In particular, while models that included both frequency and categorical predictors fared similarly, the frequency only model fared somewhat worse than the categorical predictors only models for the crossed model and the Sentence voice nested in Verb type model (> 2 standard errors away). However, this was not the case when it was compared to the Verb type nested in Sentence voice model (< 2 standard errors away), for which there was little evidence of a difference in predictive utility.

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	0
SV × VT + *SFq*	−0.8	0.8	0
SV × *VT* + SFq	0.0	0.0	0
SFq	−11.2	5.3	0
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + *SFq*	−0.9	0.8	0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	0
SFq	−11.3	5.3	0
VT + SV ∈ SE + SV ∈ OE + *SFq*	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE	−0.1	1.0	0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	0
SFq	−11.6	5.3	0

Table 20: Spillover word RTs additive model LOO comparisons (Experiment 1)

Additive comparisons In the additive comparisons for the spillover word, we found a similar pattern as before: adding frequency to a model using only categorical predictors did not affect its predictive utility, while adding categorical predictors to a model using only frequency did improve its predictability (> 2 standard errors away).

Question responses

Model	elpd_diff	se_diff	n prob. obs
SV × VT × QV	0.0	0.0	0
SV × VT × QV + SFq	−2.4	1.3	0
SFq	−23.4	8.7	0

Table 21: Question accuracy predictive crossed model LOO comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + SFq	−2.6	1.0	0
SFq	−23.0	8.7	0

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

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE + SFq	−1.7	1.2	0
SFq	−26.3	9.0	0

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

Predictive comparisons For the predictive comparisons of the question accuracy models, we found similar results to the comparisons of the models of the spillover word RTs. In particular, while models with categorical predictors performed best numerically, models with categorical predictors and Sentence voice frequency fared similarly to these. In contrast, models that used only the frequency predictors showed lower predictive utility (> 2–3 standard errors away from the best model).

Model	elpd_diff	se_diff	n prob. obs
SV × VT × QV	0.0	0.0	0
SV × VT × QV + *SFq*	−1.1	0.7	0
SV × *VT* × *QV* + SFq	0.0	0.0	0
SFq	−19.1	6.7	0
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE + *SFq*	−1.1	0.6	0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	0
SFq	−19.3	6.9	0
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + *SFq*	−0.7	0.6	0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	0
SFq	−19.1	6.9	0

Table 24: Question accuracy additive model LOO comparisons (Experiment 1)

Additive comparisons Across the different additive model comparisons for question accuracy, we found similar results to those found for the additive comparisons on the RT measures. Adding the frequency predictor to a model without it produced a numerically worse model that was nevertheless comparable in terms of its predictive utility. However, adding the categorical predictors to the frequency only model resulted in a model with greater predictive utility (> 2 standard errors away from the best model).

2.5 *Discussion*

Experiment 1 revealed two key findings:

1. **Alignment between thematic role and grammatical function facilitates processing and comprehension**: Subject experiencer verbs generally led to faster RTs and better comprehension than object experiencer verbs in active voice and subject experiencer verbs in passive voice. Object experiencer verbs generally led to faster RTs and better comprehension than subject experiencer verbs in passive voice and object experiencer verbs in active voice.

2. **Alignment effects cannot be reduced to frequency**: The alignment effects generally remained reliable even when we added verb-specific frequencies of occurring in a given voice as a predictor. Moreover, predictive modeling showed that adding frequency did not yield any improvements in predictive accuracy above and beyond categorical predictors. The reverse was not true: the addition of the categorical predictors of Verb type and Sentence voice provided significant improvement in a model's predictive utility above and beyond frequency.

For spillover word RTs and question accuracy, we found evidence of a crossover interaction, with all resolutions of the interaction between Sentence voice and Verb type showing reliable effects. The directionality of these effects is consistent with the thematic-grammatical alignment hypothesis. 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, and did not consistently control for effects of animacy, definiteness, and linear order. Our study fills these gaps.

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 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 evidence for any 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 found. 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.

In our data, as participants showed evidence of easier processing 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 the thematic-grammatical alignment hypothesis than with an experiencer-first preference.[17] 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 are thought to 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 stimuli, which had two animate nouns, found no evidence for it.

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

[17] 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.

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, and is entirely compatible with the GUPPy framework, as grammatical constraints certainly vary across languages.

We examined several possible psycholinguistic models that have been proposed in order to capture alignment effects like those we found. We believe the constraint-based approach of Wagers & Pendleton (2016) and Wagers et al. (2018) is to be favored over comprehension-specific accounts like the eADM. Though our results are indeed from comprehension (and are thus compatible with either a version of the eADM that incorporates a grammatical function hierarchy or a constraint-based approach), there are reasons both empirical and theoretical to favor the constraint-based approach. Empirically, several production studies found evidence of alignment effects with experiencer verbs consistent with the comprehension effects we found (Altmann et al. 2004; Altmann & Kemper 2006; Do & Kaiser 2019, 2022; Ferreira 1994; Gennari & MacDonald 2009), though we note that all such studies would also be compatible with an experiencer-first preference considered on their own. Nevertheless, the picture that emerges when taking our results in combination these previous studies' results leads us to favor the theory that explains both kinds of effects as arising from a single underlying source. Disentangling the effects of thematic-grammatical alignment from linear order in production like we have attempted to do in comprehension should be enlightening regarding the nature of the link between the production and comprehension systems.

We considered two frameworks that explicitly account for frequency effects: PDC, which posits a central role for frequency in comprehension (Gennari & MacDonald 2009; MacDonald 2013); and GUPPy. To evaluate their contrasting predictions, we compared models fit using only grammatical predictors, to models using only frequency predictors, to models using both. This allowed us to determine whether frequency predictors were better predictors (or at least sufficiently useful to justify their inclusion), when compared to grammatical predictors. We found fairly consistently that adding the grammatical predictors to a frequency-only model increased predictive utility. However, adding frequency to a model that used only grammatical predictors did not increase predictive utility. This is most consistent with frequency having a relationship to comprehension that is less direct and weaker than the relationship between comprehension and grammatical factors like Verb type and Sentence voice. We thus take the results of these comparisons to support GUPPy over PDC.

3. Experiment 2

The results of experiment 1 provide evidence for the importance of thematic-grammatical alignment in on-line processing and comprehension. 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). In addition, it requires participants not to reveal directly any effects of processing or comprehension difficulty, but instead to simply assess how willing they are to accept a sentence as grammatical.

This latter difference is perhaps the most important part of adopting this measure for evaluating the general feasibility of GUPPy: we proposed that the effects we found with experiencer verbs came from grammatical markedness in the form of misaligned linking configurations. Because GUPPy posits that the origin of comprehension difficulties can be found largely in the grammar, comprehenders should find misaligned configurations not only harder to process and understand, but also more grammatically marked. Thus, participants should respond more often that grammatical but marked configurations are ungrammatical, especially when under time pressure. It is for this reason that speeded grammaticality provides an ideal way of examining the predictions of GUPPy in greater detail.

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 was 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 "true" and "false"). 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 justified by the fixed effects design, except for the additive model comparisons. We also conducted frequentist analyses using R's `lme4` and `lmerTest` packages, which showed the same pattern of effects; these are included in our supplementary materials. As for experiment 1, we report estimates, estimated error, 95% CIs, and p_{MCMC} values for each effect. In addition, 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. For reasons of space, we leave a full presentation of these results to Appendix G, but present the effects found in the interim summary (section 3.2.3) in brief.

3.2.1 Voice and Verb type models

Figure 8 shows the percent of "grammatical" responses by condition.

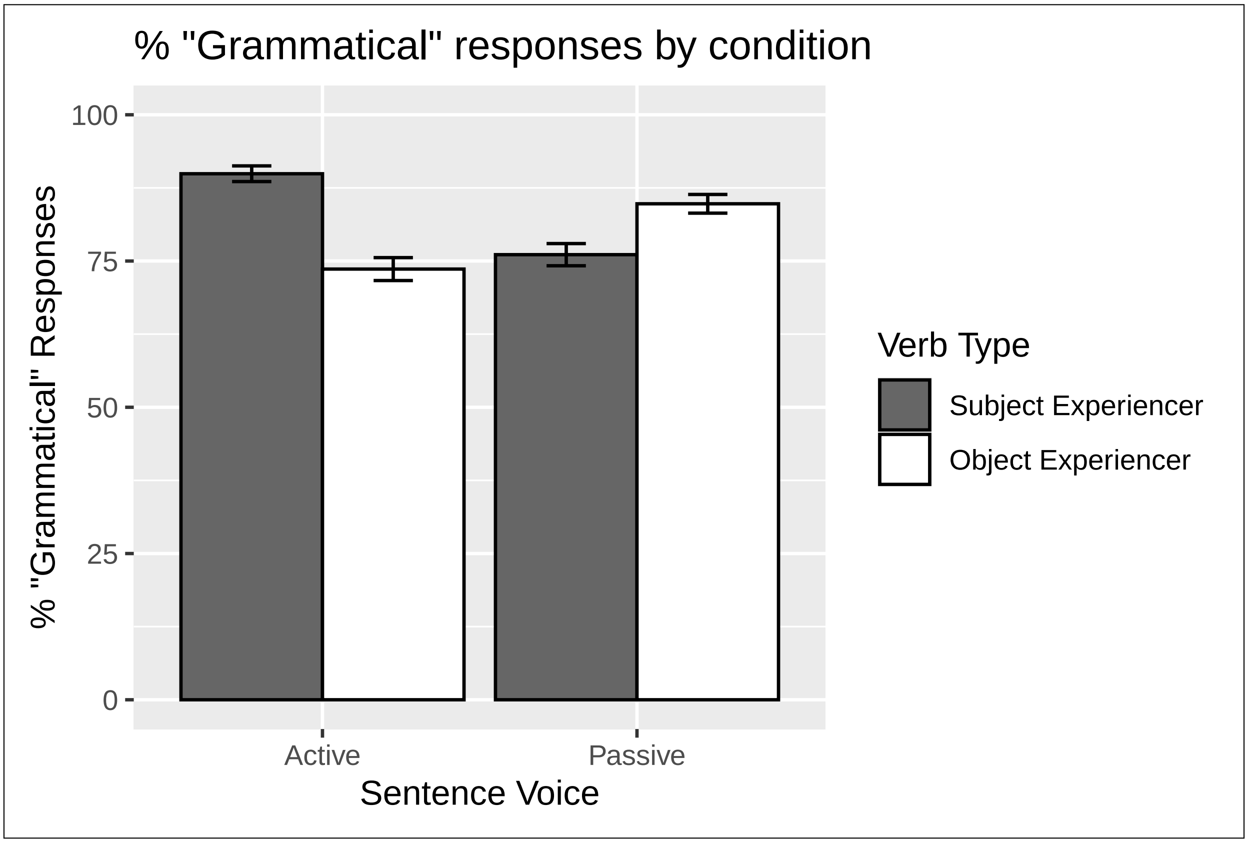


Figure 8: Percent "grammatical" responses by condition (Experiment 2)

Table 25 shows the results of the model crossing Sentence voice and Verb type.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	2.04	0.20	1.67	2.44	–
Verb type	−0.38	0.18	−0.75	−0.02	0.02
Sentence voice	0.17	0.19	−0.21	0.53	0.19
Verb type × Sentence voice	−1.86	0.40	−2.65	−1.08	0.00

Table 25: Crossed model for Response without frequency (Experiment 2)

This model found evidence for a main effect of Verb type (95% CI: [−0.75, −0.02]; p_{MCMC} = 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]; $p_{MCMC} = 0$).

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	p_{MCMC}
Verb type (active)	-1.44	0.32	-2.08	-0.83	0.00
Verb type (passive)	0.62	0.24	0.14	1.09	0.01

Table 26: Verb type nested in Sentence voice model for Response without frequency (Experiment 2)

This model nesting Verb type in Sentence voice revealed evidence for Verb type in both active (95% CI: $[-2.08, -0.83]$; $p_{MCMC} = 0.00$) and passive sentences (95% CI: $[0.14, 1.09]$; $p_{MCMC} = 0.01$), 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	p_{MCMC}
Sentence voice (subjexp)	1.16	0.29	0.60	1.73	0.00
Sentence voice (objexp)	-0.86	0.28	-1.42	-0.33	< 0.00

Table 27: Sentence voice nested in Verb type model for Response without frequency (Experiment 2)

The model nesting Sentence voice in Verb type similarly revealed evidence for an effect of Sentence voice with both subject experiencer (95% CI: $[0.60, 1.73]$; $p_{MCMC} = 0.00$) and object experiencer verbs (95% CI: $[-1.42, -0.33]$; $p_{MCMC} < 0.00$), 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.

3.2.2 Voice frequency model

In experiment 1, models using frequency only showed evidence for effects, but generally had less predictive utility than models using categorical predictors (and models using both kinds of predictors). 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 had the higher predictive utility according to LOO cross-validation and WAIC.

Figure 9 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.

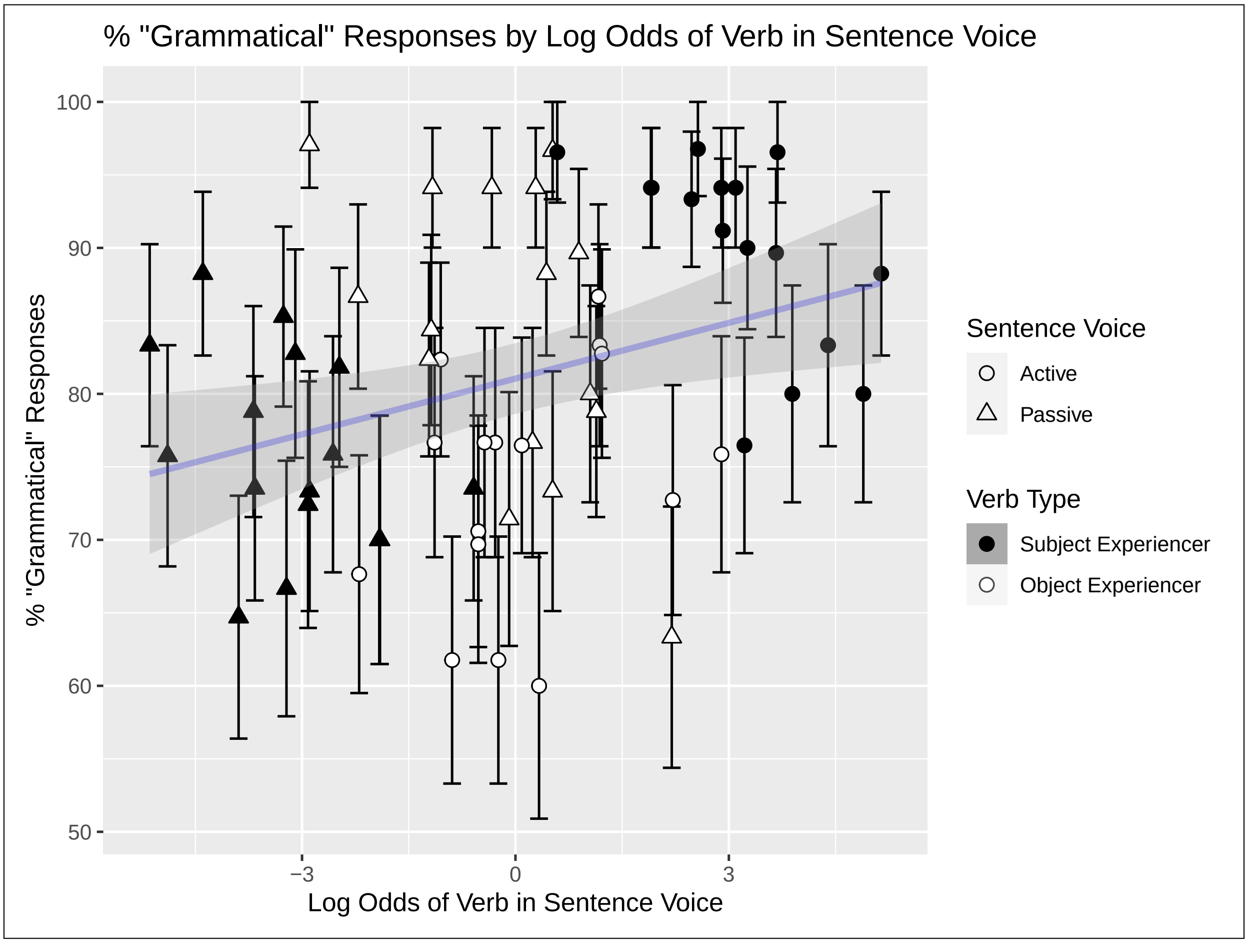


Figure 9: Regression of % "grammatical" responses on frequency (Experiment 2)

Table 28 shows the results of the logistic regression of Response on frequency. There was evidence for an effect of frequency (95% CI: [0.02, 0.18]; $p_{MCMC} = 0.01$), reflecting the fact that participants were more likely to judge a sentence as grammatical when the verb occurred in the voice it did in the sentence more frequently.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	1.89	0.18	1.55	2.26	–
Sentence voice log odds	0.10	0.04	0.02	0.18	0.01

Table 28: Frequency model for Response (Experiment 2)

3.2.3 *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 evidence for 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.

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
Frequency	Sentence voice frequency	higher frequency	↔	more “grammatical”
Both	Verb type	objexp	<	subjexp
	Sentence voice	passive	$<^{\mathrm{m}}$	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)

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

3.2.4 *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 both LOO and WAIC; we present the results of the LOO comparisons here. The WAIC results were very similar, and can be found in Appendix H.

Model	elpd_diff	se_diff	n prob. obs
SV × VT + SFq	0.0	0.0	0
SV × VT	−0.5	1.7	0
SFq	−29.9	7.7	0

Table 30: Response predictive crossed model LOO comparisons (Experiment 2)

Model	elpd_diff	se_diff	n prob. obs
SV + VT ∈ A + VT ∈ P + SFq	0.0	0.0	0
SV + VT ∈ A + VT ∈ P	−1.8	2.1	0
SFq	−31.8	8.0	0

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

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE + SFq	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE	−1.4	2.0	0
SFq	−30.7	8.2	0

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

Predictive comparisons The results of the predictive model comparisons differed slightly between experiments 1 and 2. In experiment 1, we consistently found that the numerically best models were those that did not include frequency, though they did not differ in predictive utility from the models that also included frequency. However, in experiment 2, the numerically best models were the ones that included both frequency and grammatical predictors. Nevertheless, their predictive utility was not substantially different from the models that did not include frequency.

However, like experiment 1, we found that the models for experiment 2 that used only frequency had lower predictive utility than the models that had grammatical predictors (> 3–4 standard errors away from the best model). We would thus consider the inclusion of categorical predictors to improve the predictive utility of our models, while the inclusion of frequency has little effect.

Model	elpd_diff	se_diff	n prob. obs
SV × VT + *SFq*	0.0	0.0	0
SV × VT	−1.2	1.2	0
SV × *VT* + SFq	0.0	0.0	0
SFq	−23.4	6.3	0
SV + VT ∈ A + VT ∈ P + *SFq*	0.0	0.0	0
SV + VT ∈ A + VT ∈ P	−2.5	1.6	0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	0
SFq	−24.0	6.7	0
VT + SV ∈ SE + SV ∈ OE + *SFq*	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE	−1.8	1.5	0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	0
SFq	−24.0	6.7	0

Table 33: Response additive model LOO comparisons (Experiment 2)

Additive comparisons Consistent with the pattern found in the predictive comparisons, the additive comparisons found that the models including both grammatical and frequency predictors were numerically best. However, adding frequency to models with grammatical predictors produced no difference in predictive ability (all eldp_diffs < 4), while adding the grammatical predictors to the frequency only model always produced improved predictive ability (> 3–4 standard errors away from the best model).

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. We also considered how greater experience with a particular verb in a particular voice might influence this behavior: are more frequent verb-voice combinations judged less often as ungrammatical, as predicted by frequency-based accounts?

Unlike in experiment 1, the models that numerically fared best in comparisons were consistently the models that used both grammatical and frequency-based predictors. However, while adding the grammatical predictors to the frequency only model improved predictive utility, adding frequency to the models using grammatical predictors did not noticeably improve their predictive utility. Thus, we conclude that there is not sufficient evidence to decide between the models that used only grammatical predictors and the models that used both kinds of predictors, but there is evidence against using a frequency-only model to account for experiment 2's results. This finding is, like the results of experiment 1, more consistent with GUPPy than with an alternative like PDC.

The findings of experiment 2 are not only consistent with GUPPy—like the results of experiment 1—but lend further support to it. In particular, while experiment 1 found that grammatical predictors did better than frequency at predicting processing and comprehension behavior, it provided no direct evidence that the source of difficulty associated with misaligned configurations was related to grammatical markedness. It may have been the case that such results arose specifically because participants were reading for comprehension purposes, and the grammatical factors could have been active only in comprehension (as predicted by an approach like the eADM (Bornkessel & Schlesewsky 2006; Bornkessel-Schlesewsky & Schlesewsky 2009)).

In contrast, experiment 2 showed that when participants made speeded judgments of grammaticality, a similar pattern of effects emerged, with conditions that were more difficult to process and comprehend in experiment 1 leading to fewer "grammatical" responses. While it is clear that responses such as these never truly reflect grammaticality alone, the results are suggestive of an underlying grammatical markedness associated with the conditions that proved more difficult to understand and comprehend. Thus, these results, together with the production results we reviewed, support a view where producers' preferences and comprehenders' strategies are tightly linked to grammatical representations.

4. General discussion

We conducted one self-paced reading experiment and one speeded grammaticality judgment experiment to test the predictions of the thematic-grammatical alignment hypothesis. The key prediction of this hypothesis is that language production and comprehension are facilitated when thematic role and grammatical function hierarchies are aligned. In both our experiments we found evidence in support of this hypothesis: aligned configurations (e.g., passive object experiencer verbs, active subject experiencer verbs) were easier to process, more accurately comprehended, and rated more acceptable than were misaligned configurations (e.g., active object experiencer verbs, passive subject experiencer verbs). Additional Bayesian analyses showed that this effect could not be reduced to verb-specific frequencies of occurrence in active or passive voice.

We consider our results to support the thematic-grammatical alignment hypothesis in a more direct way than previous work. While previous psycholinguistic literature has found effects consistent with alignment in general, such effects were not always consistent across and within studies, and were compatible with a number of competing explanations. Our study controlled for factors such as animacy, definiteness, and linear order, and found clear effects of thematic-grammatical alignment on processing, comprehension, and grammaticality judgments. No prior study we are aware of has controlled for all of these factors in a within-subjects design; our results fill this gap in the prior literature, and thus provide clear support for a role of thematic-grammatical alignment in processing. While our aim was to control for potential confounds of animacy, definiteness, identity/similarity of thematic roles, linear order, and voice, we would like to emphasize that our results are not inconsistent with the substantial body of work showing effects of these factors; alignment occurs on many hierarchies simultaneously, and these hierarchies may well include ones based on animacy, linear order, and definiteness, which may interact in interesting ways with the thematic role and grammatical function hierarchies we focused on.

Our results raise the question of how exactly thematic-grammatical alignment influences comprehension processes. When our results are considered in the context of the previous literature, which is consistent with the thematic-grammatical alignment hypothesis (most clearly in the work of Ferreira (1994) and Do & Kaiser (2019, 2022)), a fairly consistent picture emerges: sentences that exhibit thematic-grammatical alignment 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 cognitive constraints operate in tandem to influence both production and comprehension directly, where frequency is related to the resulting production preferences (Gennari & MacDonald 2009; MacDonald 2013), but has a less direct relationship to comprehension than do grammatical and cognitive constraints. This general framework proposes a direct role for grammatical constraints in comprehension, and builds on recent constraint-based proposals (notably Wagers & Pendleton 2016 and Wagers et al. 2018). We called this the Grammar Underlies Production and Processing (plus frequency) framework (GUPPy). GUPPy recalls other work that has argued for the central importance of grammatical representations to language use (e.g., Hirsch & Hartman 2006; Momma & Phillips 2018). The advantage of GUPPy is that it treats all effects of thematic-grammatical alignment, across both production and comprehension, as having a single underlying source in the grammatical constraints that govern linking behavior.

There are several caveats worth noting. First is that we are unaware of any production study that has deconfounded linear order and grammatical role the way our study did. This would be important for establishing a role for thematic-grammatical alignment in production, as opposed to an experiencer-first bias. Second, the work here is based on English, and extending the present design to languages other than English is critical.

Additionally, we compared models that used grammatical predictors (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 them. We take these results to support GUPPy's approach to integrating distributional frequencies with a causal model of sentence production and processing over PDC's. In particular, the fact that grammatical markedness had more consistent and clear effects on processing and comprehension than frequency is most consistent with a causal model where grammatical constraints affect processing more directly than frequency. Nevertheless, in models without predictors related to grammatical markedness, we found effects of frequency, which is consistent with frequency being implicated in processing in a less direct manner.

This latter observation raises important questions about how frequency distributions arise, and how they influence comprehension processes. We note that approaches where frequency is a predictor must necessarily grapple with the question of the source of various distributional patterns. PDC answers this in one particular way, by treating distributional frequencies as arising from production preferences that are imposed by cognitive constraints. This is a markéd improvement over treating frequency as reflecting random variation that becomes grammatically reified over time. However, 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 would differ from the PDC and GUPPy 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 cognitive, pragmatic, and discourse factors. Thus, we expect a relationship between frequency and (un)marked structures, but an indirect one at best. This explains why we found weaker evidence for effects of frequency in our experiments than might be predicted by an approach like GUPPy which posits a single underlying source for production and processing effects of thematic-grammatical alignment, as well as why the results of our corpus study did not perfectly mirror the markedness of various 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. It is for this reason that frequency reflects only noisily the grammatical constraints which underlie production and processing, as pragmatic, cognitive, and discourse constraints play roles in determining frequency as well.

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.

5. Conclusion

We conducted two experiments examining how the linking properties of experiencer verbs influence sentence processing, comprehension, and judgments of grammaticality. 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 the thematic-grammatical alignment hypothesis, and show no evidence for an experiencer-first preference in the processing of English.

In our view, one of the more important methodological contributions of our study is to show the importance of considering both models based on grammatical constraints and models based on frequency when comparing approaches that differ in the relative importance they assign to each kind of factor. In particular, if we had only considered frequency-based models, we would have concluded that frequency plays an important role in comprehension. While we do not deny this conclusion, our approach of comparing different models showed that models using grammatical predictors were better predictors than those that used only frequency. This suggests a much more limited role for frequency.

Nevertheless, the effects of frequency we found when omitting grammatical predictors are consistent with GUPPy, which posits that the reason for the relationship between frequency and comprehension is primarily due to the presence of an underlying contributor – grammatical constraints – to comprehension. We hope that in the future comparing results according to different causal models, as we have here, will be more common in psycholinguistic research, as this should prove especially useful in disentangling the predictions and empirical adequacy of competing frameworks.

References

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.

Baird, R., & Koslick, J. D. (1974). Recall of grammatical relations within clause-containing sentences. *Journal of Psycholinguistic Research*, *3*(2), 165–171.

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.

Bennis, H. (2004). Unergative adjectives and psych verbs. In A. Alexiadou, E. Anagnostopoulou, & M. Everaert (Eds.) *The Unaccusativity Puzzle: Explorations of the Syntax-Lexicon Interface*, (pp. 129–137). Oxford, UK: Oxford University Press.

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

van Bergen, G. (2011). *Who's First and What's Next: Animacy and Word Order Variation in Dutch Language Production*. Ph.D. thesis, Radboud Universiteit Nijmegen.

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.

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.) *Handbook of Psycholinguistics*, (pp. 945–984). San Diego, CA: Academic Press.

Bondaruk, A., Rozwadowska, B., & Witkowski, W. (2017). Passivisation of Polish object experiencer verbs vs. the unaccusativity hypothesis (part 2). *Studies in Polish Linguistics*, *12*(3), 123–144.

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.

Caplan, D., & Waters, G. S. (1992). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(1), 77–94.

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

Cheung, C. C.-H., & Larson, R. (2015). Psych verbs in English and Mandarin. *Natural Language & Linguistic Theory*, *33*, 127–189.

Chung, S. (2012). On reaching agreement late. In A. Beltrama, T. Chatzikonstantinou, J. L. Lee, M. Pham, & D. Rak (Eds.) *Proceedings of the Forty-eighth Annual Meeting of the Chicago Linguistic Society*, (pp. 161–190). Chicago: Chicago Linguistic Society.

Clothier-Goldschmidt, S. (2015). *The Distribution and Processing of Referential Expressions: Evidence from English and Chamorro*. Master's thesis, University of California, Santa Cruz.

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

Cuervo, M. C. (2003). A control-vs-raising theory of dative experiencers. In A. T. Pérez-Leroux, & Y. Roberge (Eds.) *Romance Linguistics: Theory and Acquisition: Selected Papers from the 32nd Linguistic Symposium on Romance Languages (LSRL), Toronto, April 2002*, vol. 244 of *Current Issues in Linguistic Theory*, (pp. 111–130). Amsterdam/Philadelphia: John Benjamins.

——— (2010). Some dative subjects are born, some are made. In C. Borgonovo, M. Español Echevarría, & P. Prévost (Eds.) *Selected Proceedings of the 12th Hispanic Linguistics Symposium*, (pp. 26–37). Somerville, MA: Cascadilla Press.

——— (2020). Datives as applicatives. In A. Pineda, & J. Mateu (Eds.) *Dative Constructions in Romance and Beyond*, (pp. 1–39). Berlin: Language Science Press.

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.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B*, *369*(1634), 20120394.

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*. Amsterdam: John Benjamins.

Drummond, A. (2011). *IbexFarm (Version 0.3.9) [Software]*. Now defunct; formerly available online at https://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*(2), 164–203.

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

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 203–218.

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

Friederici, A. D., Steinhauer, K., Mecklinger, A., & Meyer, M. (1998). Working memory constraints on syntactic ambiguity resolution as revealed by electrical brain responses. *Biological Psychology*, *47*(3), 193–221.

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.

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, *65*(2), 141–176.

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

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

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411–1423.

——— (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, *51*(1), 97–114.

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.

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.

He, W., & Chen, B. (2013). The role of animacy in Chinese relative clause processing. *Acta Psychologica*, *144*(1), 145–153.

Hirsch, C., & Hartman, J. (2006). Some (*wh*-)questions concerning passive interactions. In A. Belletti, E. Bennati, C. Chesi, E. Di Domenico, & I. Ferrari (Eds.) *Proceedings of the Conference on Generative Approaches to Language Acquisition*, (pp. 256–268). Cambridge, UK: Cambridge Scholars Press.

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.

Hirsch, N. (2018). *German Psych Verbs – Insights from a Decompositional Perspective*. Ph.D. thesis, Humboldt-Universität zu Berlin.

Hornstein, N., & Motomura, M. (2002). Psych verbs, theta roles and reconstruction. In *Proceedings of the 2002 Linguistic Society of Korea International Summer Conference*, vol. 2, (pp. 39–58). Seoul: Linguistic Society of Korea.

Hsiao, Y., & MacDonald, M. C. (2016). Production predicts comprehension: Animacy effects in Mandarin relative clause processing. *Journal of Memory and Language*, *89*, 87–109.

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.

Kaiser, E., Holsinger, E., & Li, D. C.-H. (2011). It's not the words, but how you say them: Effects of referential predictability on the production of names and pronouns. In K. van Deemter, A. Gatt, R. van Gompel, & E. Krahmer (Eds.) *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the Gap between Computational, Empirical & Theoretical Approaches to Reference*. Austin, TX: Cognitive Science Society. Published online at https://pre2011.uvt.nl/pdf/kaiser.pdf.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *8*(1), 63–99.

Keshev, M., & Meltzer-Asscher, A. (2021). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. *Cognitive Psychology*, *124*, 101359.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602.

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. (2002). A typology of psych passives. In M. Hirotani (Ed.) *Proceedings of the North East Linguistic Society 32*, vol. 1, (pp. 271–286). Amherst, MA: Graduate Linguistics Students Association.

——— (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.

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. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney (Ed.) *The Emergence of Language*, vol. 16 of *Carnegie Mellon Symposia on Cognition*, (pp. 177–196). Mahwah, NJ: Lawrence Erlbaum Associates.

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

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler, & M. A. Gernsbacher (Eds.) *Handbook of Psycholinguistics*, (pp. 581–611). London: Academic Press, 2 ed.

MacDonald, M. C., & Thornton, R. (2009). When language comprehension reflects production constraints: Resolving ambiguities with the help of past experience. *Memory & Cognition*, *37*(8), 1177–1186.

MacWhinney, B., & Pléh, C. (1988). The processing of restrictive relative clauses in Hungarian. *Cognition*, *29*(2), 95–141.

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory and Language*, *47*(1), 50–68.

——— (2006). Animacy in processing relative clauses: The hikers that rocks crush. *Journal of Memory and Language*, *54*(4), 466–490.

Malaia, E. (2009). Ontological representations of event structure across languages: The case of psych verbs. In *Proceedings of the 18th International Congress of Linguists*, (pp. 2727–2740). Seoul: Permanent International Committee of Linguists.

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.

McElreath, R. (Unpublished). Statistical rethinking^2: A Bayesian course with examples in R and Stan. Cited version compiled 8 December 2019.

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.

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 `https://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.

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.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*, vol. 2 of *Rutgers University Center for Cognitive Science Technical Reports*. Piscataway, NJ: Rutgers University Center for Cognitive Science.

Pylkkänen, L. (2000). On stativity and causation. In C. L. Tenny, & J. Pustejovsky (Eds.) *Events as Grammatical Objects: The Converging Perspectives of Lexical Semantics and Syntax*, vol. 100 of *CSLI Lecture Notes*, (pp. 417–444). Stanford: Center for the Study of Language and Information Publications.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.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.

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). Uncertainty in Bayesian Leave-One-Out cross-validation based model comparison. Available at https://arxiv.org/abs/2008.10296.

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

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

Staub, A., Dillon, B., & Clifton, C., Jr. (2017). The matrix verb as a source of comprehension difficulty in object relative clauses. *Cognitive Science*, *41*(S6), 1353–1376.

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.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*(4), 566–585.

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

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

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.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*(1), 79–112.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594.

Waters, G. S., & Caplan, D. (1996a). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review*, *103*(4), 761–772.

——— (1996b). Processing resource capacity and the comprehension of garden path sentences. *Memory & Cognition*, *24*(3), 342–355.

Waters, G. S., Caplan, D., & Hildebrandt, N. (1987). Working memory and written sentence comprehension. In M. Coltheart (Ed.) *Attention and Performance XII: The Psychology of Reading*, (pp. 531–555). Mahwah, NJ: Lawrence Erlbaum Associates.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*(2), 250–271.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*(3), 165–209.

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

Appendix 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.

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)

Appendix B. Experiment 1: 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, we found evidence for main effects and interactions of these factors. However, in the frequency analyses, we found effects of frequency. For this reason, 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 2.4.4.

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 (16)). 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.

Appendix B.1 Critical verb

The model fit using categorical predictors as well as frequency for the critical verb revealed only evidence for a main effect of Sentence voice (95% CI: [−0.14, −0.00]; $p_{MCMC} = 0.02$), 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.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.08	0.05	5.98	6.17	–
Verb type	−0.00	0.02	−0.04	0.04	0.45
Sentence voice	−0.08	0.04	−0.14	−0.00	0.02
Sentence voice log odds	−0.01	0.01	−0.02	0.01	0.26
Verb type × Sentence voice	−0.02	0.07	−0.15	0.12	0.40

Table 34: Crossed model for critical verb with frequency (Experiment 1)

Appendix B.2 Spillover word

The model containing the categorical predictors and frequency for the spillover word revealed no evidence for any main effects. Nevertheless, there was evidence for an interaction between Verb type and Sentence voice (95% CI: [−0.38, −0.03]; $p_{MCMC} = 0.01$). This was consistent with the model fit using only categorical predictors, which revealed evidence for the same interaction; but contrasted with the model fit using only frequency, as that model revealed evidence for a main effect of frequency (see section 2.4.2).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.15	0.05	6.05	6.25	–
Verb type	0.02	0.03	−0.03	0.07	0.19
Sentence voice	−0.05	0.04	−0.14	0.03	0.11
Sentence voice log odds	−0.01	0.01	−0.03	0.01	0.16
Verb type × Sentence voice	−0.20	0.09	−0.38	−0.03	0.01

Table 35: Crossed model for spillover word with frequency (Experiment 1)

We note, however, 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 3 showed, the 95% CI for the interaction effect in the model using only the categorical predictors was [−0.39, −0.14], while this range for the interaction effect in the model that added frequency was [−0.38, 0.03] (increasing the upper bound by 0.11). 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 revealed evidence for an effect of Verb type in active sentences (95% CI: [0.02, 0.23]; $p_{MCMC} = 0.01$), 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. There was also marginal evidence for an effect of Verb type in passive sentences, which went in the opposite direction (95% CI: [−0.18, 0.02]; $p_{MCMC} = 0.06$). This latter result contrasted with the Verb type nested in Sentence voice model that did not include frequency, as that model revealed stronger evidence for an effect of Verb type in passive sentences.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Verb type (active)	0.12	0.05	0.02	0.23	0.01
Verb type (passive)	−0.08	0.05	−0.18	0.02	0.06

Table 36: Verb type nested in Sentence voice model for spillover word with frequency (Experiment 1)

The model nesting Sentence voice within Verb type found evidence only for an effect of Sentence voice in object experiencer sentences (95% CI: [−0.23, −0.08]; $p_{MCMC} = 0$). 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 evidence for 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	p_{MCMC}
Sentence voice (subjexp)	0.04	0.08	−0.12	0.20	0.29
Sentence voice (objexp)	−0.16	0.04	−0.23	−0.08	0.00

Table 37: Sentence voice nested in Verb type model for spillover word with frequency (Experiment 1)

We believe it likely that this is because subject experiencer verbs all occurred most often in the active voice (cf. figures 2–3), 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.

Appendix B.3 Question responses

The model regressing question accuracy on the categorical predictors and their interactions as well as Sentence voice frequency revealed evidence for a main effect of Sentence voice (95% CI: [−0.65, 0.05]; $p_{MCMC} = 0.04$), a main effect of Question voice (95% CI: [−0.01, 0.43]; $p_{MCMC} = 0.03$), and for an interaction between Verb type and Sentence voice (95% CI: [1.10, 2.38]; $p_{MCMC} = 0$).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	0.66	0.10	0.47	0.86	–
Verb type	−0.04	0.13	−0.30	0.22	0.37
Sentence voice	−0.30	0.18	−0.65	0.05	0.04
Question voice	0.21	0.11	−0.01	0.43	0.03
Sentence voice log odds	0.03	0.04	−0.05	0.12	0.24
Verb type × Sentence voice	1.74	0.32	1.10	2.38	0.00
Verb type × Question voice	−0.24	0.23	−0.70	0.22	0.15
Sentence voice × Question voice	0.01	0.23	−0.45	0.48	0.48
Verb type × Sentence voice × Question voice	0.36	0.38	−0.39	1.11	0.17

Table 38: 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 revealed evidence for 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	p_{MCMC}
Verb type (active)	−0.95	0.22	−1.39	−0.51	0.00
Verb type (passive)	0.86	0.21	0.46	1.28	0.00
Verb type (active) × Question voice	−0.43	0.30	−1.01	0.17	0.07
Verb type (passive) × Question voice	−0.05	0.30	−0.64	0.53	0.43

Table 39: 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	p_{MCMC}
Sentence voice (subjexp)	−1.22	0.31	−1.83	−0.62	< 0.00
Sentence voice (objexp)	0.60	0.15	0.31	0.90	0.00
Sentence voice (subjexp) × Question voice	−0.18	0.28	−0.73	0.38	0.26
Sentence voice (objexp) × Question voice	0.22	0.34	−0.44	0.90	0.25

Table 40: Sentence Voice nested in Verb type model for question accuracy with Sentence voice frequency (Experiment 1)

Appendix C. Experiment 1: Results of WAIC comparisons

This appendix presents the results of the WAIC comparisons run on the models for experiment 1 (see section 2.4.4). WAIC is defined as:

$$\text{(18)} \qquad \text{WAIC}\,(y, \Theta) = -2\left(\sum_{i} \log \frac{1}{S} \sum_{s} p\,(y_i|\Theta_s) - \sum_{i} \sigma^2_{\Theta} \log p\,(y_i|\Theta)\right)$$

where y is the measured sample values, Θ is the posterior distribution, S is the number of samples in y, and Θ_s is the s-th set of sampled parameter values in the posterior (McElreath Unpublished). A relatively lower WAIC score in a comparison between models indicates a better fit to the data. Though an absolute WAIC score is not informative, we can compare the WAIC scores of models fit to the same data to determine which has more predictive utility. However, rather than reporting the WAIC scores directly, we report the elpd_diff from the `waic` comparison function of `brms` (as we did for the LOO comparisons in the main text). This compares the models based on the elpd_waic, which is equal to $-0.5 \cdot \text{WAIC}\,(y, \Theta)$ (thereby undoing the multiplication by -2 in the definition of WAIC). Thus, models with higher WAIC scores have lower elpd_waic scores. The elpd_diff compares the difference in these elpd_waic scores relative to the model with the highest elpd_waic score; negative differences indicate a lower predictive value of the model compared to the one with the highest predictive value, just as with LOO elpd_diff scores.

We simply present the tables here, as the pattern of results was essentially identical to the LOO comparisons presented in the main text, with the exception that the WAIC comparisons found more problematic data points. For WAIC, “n prob. obs” refers to the number of points identified as having $p_{\text{WAIC}} > 0.4$.

Appendix C.1 Critical verb

Appendix C.1.1 Predictive WAIC comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	73
SV × VT + SFq	−2.4	0.8	53
SFq	−2.7	4.9	76

Table 41: Critical verb RTs predictive crossed model WAIC comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
SV + VT ∈ A + VT ∈ P	0.0	0.0	73
SV + VT ∈ A + VT ∈ P + SFq	−2.6	0.9	53
SFq	−3.7	4.9	77

Table 42: Critical verb RTs predictive Verb type nested in Sentence voice model WAIC comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	70
VT + SV ∈ SE + SV ∈ OE + SFq	−3.0	0.9	53
SFq	−5.5	5.6	72

Table 43: Critical verb RTs predictive Sentence voice nested in Verb type model WAIC comparisons (Experiment 1)

Appendix C.1.2 Additive comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	73
SV × VT + *SFq*	−0.8	0.7	70
SV × *VT* + SFq	0.0	0.0	53
SFq	−3.7	3.4	60
SV + VT ∈ A + VT ∈ P	0.0	0.0	73
SV + VT ∈ A + VT ∈ P + *SFq*	−1.0	0.7	72
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	53
SFq	−3.4	3.4	58
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	70
VT + SV ∈ SE + SV ∈ OE + *SFq*	−1.7	0.8	69
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	53
SFq	−3.5	3.4	58

Table 44: Critical verb RTs additive model WAIC comparisons (Experiment 1)

Appendix C.2 Spillover word

Appendix C.2.1 Predictive comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	90
SV × VT + SFq	−2.1	1.3	58
SFq	−18.0	8.7	97

Table 45: Spillover word RTs predictive crossed model WAIC comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
SV + VT ∈ A + VT ∈ P	0.0	0.0	94
SV + VT ∈ A + VT ∈ P + SFq	−1.6	1.6	58
SFq	−13.9	8.3	97

Table 46: Spillover word RTs predictive Verb type nested in Sentence voice model WAIC comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	84
VT + SV ∈ SE + SV ∈ OE + SFq	−0.1	1.4	58
SFq	−17.0	8.0	93

Table 47: Spillover word RTs predictive Sentence voice nested in Verb type model WAIC comparisons (Experiment 1)

Appendix C.2.2 Additive comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT	0.0	0.0	90
SV × VT + *SFq*	−0.8	0.8	89
SV × *VT* + SFq	0.0	0.0	58
SFq	−11.1	5.3	64
SV + VT ∈ A + VT ∈ P	0.0	0.0	94
SV + VT ∈ A + VT ∈ P + *SFq*	−0.8	0.8	92
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	58
SFq	−11.2	5.3	62
VT + SV ∈ SE + SV ∈ OE + *SFq*	0.0	0.0	84
VT + SV ∈ SE + SV ∈ OE	−0.3	1.0	84
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	58
SFq	−11.6	5.3	62

Table 48: Spillover word RTs additive model WAIC comparisons (Experiment 1)

Appendix C.3 Question responses

Appendix C.3.1 Predictive comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT × QV	0.0	0.0	0
SV × VT × QV + SFq	−2.1	1.3	0
SFq	−24.5	8.7	0

Table 49: Question accuracy predictive crossed model WAIC comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + SFq	−2.4	1.0	0
SFq	−24.3	8.7	0

Table 50: Question accuracy predictive Verb type nested in Sentence voice model WAIC comparisons (Experiment 1)

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE + SFq	−1.5	1.2	0
SFq	−27.5	9.0	0

Table 51: Question accuracy predictive Sentence voice nested in Verb type model WAIC comparisons (Experiment 1)

Appendix C.3.2 Additive comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT × QV	0.0	0.0	0
SV × VT × QV + *SFq*	−1.0	0.7	0
SV × *VT* × *QV* + SFq	0.0	0.0	0
SFq	−19.1	6.7	0
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0
VT + SV ∈ SE + SV ∈ OE + *SFq*	−1.0	0.6	0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	0
SFq	−19.3	6.9	0
SV + VT ∈ A + VT ∈ P	0.0	0.0	0
SV + VT ∈ A + VT ∈ P + *SFq*	−0.7	0.6	0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	0
SFq	−19.2	6.9	0

Table 52: Question accuracy additive model WAIC comparisons (Experiment 1)

Appendix D. Experiment 1: Question voice frequency models

Figure 10 shows the regression of mean question accuracy for each verb on Question voice frequency.

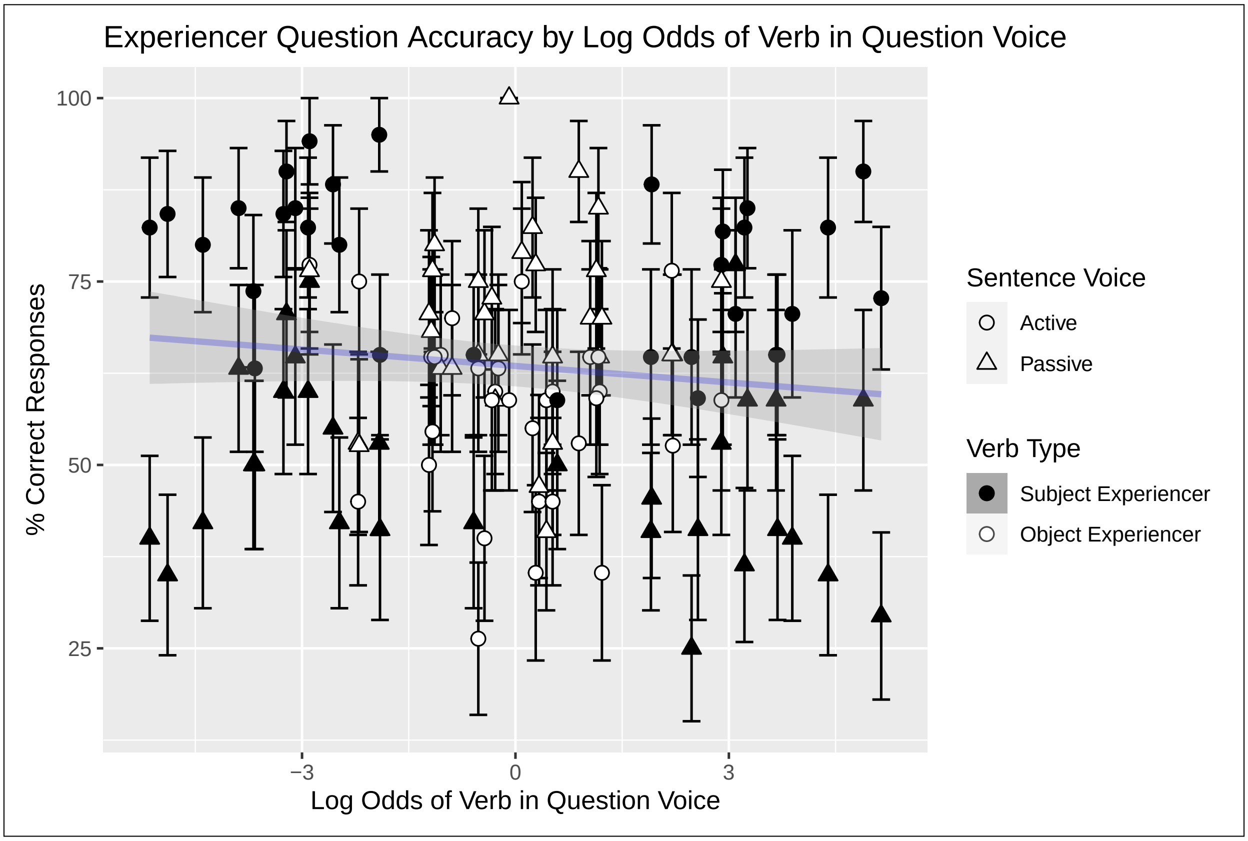


Figure 10: Regression of question accuracy on Question voice log odds (Experiment 1)

Appendix D.1 Sentence voice frequency × Question voice frequency model

The question accuracy model fit with the interaction of Sentence voice frequency and Question voice frequency revealed evidence for a main effect of Sentence voice frequency (95% CI: [0.13, 0.23]; $p_{MCMC} = 0$), as well as evidence for a main effect of Question voice frequency (95% CI: [−0.11, 0.00]; $p_{MCMC} = 0.04$), but no evidence for any interaction. The main effect of Sentence voice reflected the fact that participants answered questions accurately more often when the verb occurred in the sentence in its more frequent voice than when it occurred in its less frequent voice. However, the main effect of Question voice reflected the fact that, to some extent, participants were more likely to answer questions accurately when the verb occurred in the question in its *less* frequent voice than when it occurred in its more frequent voice.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	0.63	0.09	0.44	0.81	–
Sentence voice log odds	0.18	0.03	0.13	0.23	0.00
Question voice log odds	−0.05	0.03	−0.11	0.00	0.04
Sentence × Question voice log odds	−0.00	0.01	−0.01	0.01	0.30

Table 53: Sentence voice frequency × Question voice frequency model for question accuracy (Experiment 1)

Appendix D.2 Categorical predictors with Sentence voice frequency × Question voice frequency models

The model regressing question accuracy on the categorical predictors (and their interactions), as well as on the interaction between Sentence voice and Question voice frequency, revealed evidence for a main effect of Sentence voice (95% CI: [−0.65, 0.05]; $p_{MCMC} = 0.04$), as well as evidence for an interaction of Verb type and Sentence voice (95% CI: [1.12, 2.39]; $p_{MCMC} = 0$). The effect of Sentence voice reflected that participants answered questions about passive sentences accurately less often than questions about active sentences. These results contrasted with the results of the model fit using only the categorical predictors, which additionally found evidence for an effect of Question voice; and also contrasted with the results of the model fit using the interaction of the frequency predictors which found evidence for main effects of both Sentence voice frequency and Question voice frequency.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	0.68	0.10	0.48	0.88	–
Verb type	−0.06	0.13	−0.33	0.20	0.31
Sentence voice	−0.30	0.18	−0.65	0.05	0.04
Question voice	0.18	0.18	−0.18	0.54	0.16
Sentence voice log odds	0.04	0.05	−0.05	0.13	0.20
Question voice log odds	−0.01	0.05	−0.11	0.08	0.39
Verb type × Sentence voice	1.76	0.33	1.12	2.39	0.00
Verb type × Question voice	−0.20	0.34	−0.88	0.49	0.28
Sentence voice × Question voice	0.08	0.34	−0.57	0.75	0.41
Verb type × Sentence voice × Question voice	0.27	0.50	−0.70	1.27	0.29
Sentence × Question voice log odds	−0.00	0.01	−0.02	0.02	0.39

Table 54: Crossed model for question accuracy with Sentence and Question voice frequency (Experiment 1)

We resolved the interaction between Verb type and Sentence voice by both Verb type and Sentence voice. This revealed the same pattern of effects found in the nested models fit using only the categorical predictors, so we do not discuss them further.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Verb type (active)	−0.98	0.23	−1.42	−0.53	0.00
Verb type (passive)	0.85	0.21	0.44	1.27	0.00
Verb type (active) × Question voice	−0.32	0.43	−1.16	0.51	0.23
Verb type (passive) × Question voice	−0.02	0.42	−0.84	0.82	0.48

Table 55: Verb type nested in Sentence voice model for question accuracy with Sentence and Question voice frequency (Experiment 1)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Sentence voice (subjexp)	−1.24	0.32	−1.85	−0.62	0.00
Sentence voice (objexp)	0.61	0.15	0.30	0.91	< 0.00
Sentence voice (subjexp) × Question voice	−0.15	0.48	−1.09	0.80	0.38
Sentence voice (objexp) × Question voice	0.23	0.34	−0.44	0.90	0.25

Table 56: Sentence voice nested in Verb type model for question accuracy with Sentence and Question voice frequency (Experiment 1)

Appendix D.3 Model comparisons

Appendix D.3.1 Predictive comparisons

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV × VT × QV	0.0	0.0	0.0	0.0	0; 0
SV × VT × QV + SFq	−2.1	1.3	−2.4	1.3	0; 0
SV × VT × QV + SFq × QFq	−3.3	2.9	−4.5	2.9	15; 0
SFq × QFq	−22.4	8.5	−22.0	8.5	5; 0
SFq	−24.5	8.7	−23.4	8.7	0; 0

Table 57: Question accuracy predictive crossed model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV + VT ∈ A + VT ∈ P	0.0	0.0	0.0	0.0	0; 0
SV + VT ∈ A + VT ∈ P + SFq	−2.4	1.0	−2.6	1.0	0; 0
SV + VT ∈ A + VT ∈ P + SFq × QFq	−3.5	2.9	−4.7	2.9	14; 0
SFq × QFq	−22.2	8.5	−21.7	8.5	5; 0
SFq	−24.3	8.7	−23.0	8.7	0; 0

Table 58: Question accuracy predictive Verb type nested in Sentence voice model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0.0	0.0	0; 0
VT + SV ∈ SE + SV ∈ OE + SFq	−1.5	1.2	−1.7	1.2	0; 0
VT + SV ∈ SE + SV ∈ OE + SFq × QFq	−4.1	2.8	−5.3	2.8	14; 0
SFq × QFq	−25.4	8.8	−25.0	8.8	5; 0
SFq	−27.5	9.0	−26.3	9.0	0; 0

Table 59: Question accuracy predictive Sentence voice nested in Verb type model comparisons (Experiment 1)

For the predictive comparisons of the question accuracy models, we found similar results to the comparisons of the models of the RT measures. In particular, while models with categorical predictors performed best numerically, models with categorical predictors and Sentence voice frequency and/or Question voice frequency predictors fared similarly to these. In constrast, models that used only the frequency predictors showed lower predictive utility (> 2–3 standard errors away from the best model).

Appendix D.4 Additive comparisons

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SFq	0.0	0.0	0.0	0.0	0; 0
SFq × *QFq*	0.0	2.2	0.0	2.2	0; 0
SV × VT × QV	0.0	0.0	0.0	0.0	0; 0
SV × VT × QV + *SFq* × *QFq*	−2.7	0.8	−2.7	0.8	0; 0
SV × *VT* × *QV* + SFq × QFq	0.0	0.0	0.0	0.0	5; 0
SFq × QFq	−17.1	6.4	−17.0	6.4	6; 0
SV × VT × QV + SFq	0.0	0.0	0.0	0.0	0; 0
SV × VT × QV + SFq × *QFq*	−1.5	0.4	−1.5	0.4	1; 0
SV × *VT* × *QV* + SFq × *QFq*	0.0	0.0	0.0	0.0	0; 0
SFq	−17.2	6.7	−17.2	6.7	0; 0

Table 60: Question accuracy additive crossed model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0.0	0.0	0; 0
VT + SV ∈ SE + SV ∈ OE + *SFq* × *QFq*	−2.5	0.7	−2.6	0.7	0; 0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq × QFq	0.0	0.0	0.0	0.0	5; 0
SFq × QFq	−17.3	6.6	−17.2	6.6	5; 0
VT + SV ∈ SE + SV ∈ OE + SFq	0.0	0.0	0.0	0.0	0; 0
VT + SV ∈ SE + SV ∈ OE + SFq × *QFq*	−1.9	0.4	−2.0	0.4	1; 0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq × *QFq*	0.0	0.0	0.0	0.0	0; 0
SFq	−17.2	6.9	−17.1	6.9	0; 0

Table 61: Question accuracy additive Verb type in Sentence voice model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV + VT ∈ A + VT ∈ P	0.0	0.0	0.0	0.0	0; 0
SV + VT ∈ A + VT ∈ P + *SFq* × *QFq*	−2.5	0.6	−2.5	0.6	0; 0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq × QFq	0.0	0.0	0.0	0.0	5; 0
SFq × QFq	−16.9	6.6	−16.8	6.6	4; 0
SV + VT ∈ A + VT ∈ P + SFq	0.0	0.0	0.0	0.0	0; 0
SV + VT ∈ A + VT ∈ P + SFq × *QFq*	−1.5	0.3	−1.6	0.3	0; 0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq × *QFq*	0.0	0.0	0.0	0.0	0; 0
SFq	−17.5	6.9	−17.5	6.9	0; 0

Table 62: Question accuracy additive Sentence voice in Verb type model comparisons (Experiment 1)

Across the different additive model comparisons for question accuracy, we found similar results to those found for the additive comparisons on the RT measures. Adding the frequency predictor(s) to a model without them produced a numerically worse model that was nevertheless comparable in terms of its predictive utility. However, adding the categorical predictors to the frequency only models consistently resulted in a model with greater predictive utility (> 2 standard errors away from the best model).

Appendix E. Experiment 1: Region-based analyses

In addition to the analysis of RTs of individual words we discuss in the main text, we also considered log reading times at a region level that added reading times from individual words together (prior to applying the log function). The reason for using a region-based measure in addition to a word-by-word one was to create comparable analysis regions across active and passive conditions, which contained different numbers of words (the auxiliary and *by* occur in passives but not actives). An example of the region breaks we used is the following, with regions separated by a forward slash:

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

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

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

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

Appendix E.1 Sentence voice and Verb type models

Figure 11 displays region-based log RTs by Sentence voice and Verb type.

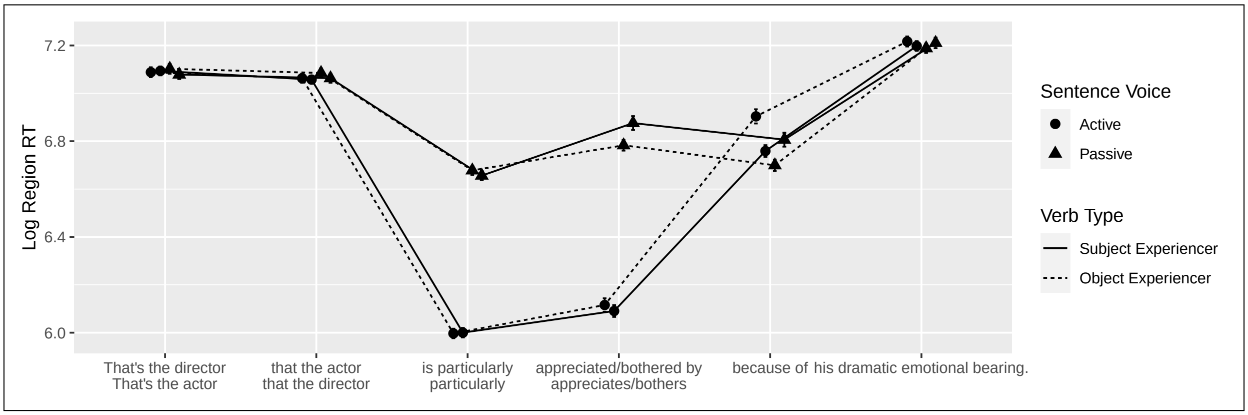


Figure 11: Log RTs by region (Experiment 1)

Appendix E.1.1 Critical region

The model summarized in table 63 shows evidence for a main effect of Sentence voice at the critical region (95% CI: [0.68, 0.77]; $p_{MCMC} = 0$); this was expected in the region-based analysis because active sentences contained one word in the critical region (the critical verb), while passive sentences contained more than one word (the critical verb + *by*). This effect consistently appears in the critical region in our region-based analyses, and we do not discuss it further.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.46	0.05	6.37	6.56	–
Verb type	−0.03	0.02	−0.08	0.01	0.95
Sentence voice	0.73	0.02	0.68	0.77	0.00
Verb type × Sentence voice	−0.12	0.05	−0.21	−0.02	0.01

Table 63: Crossed model for critical region without frequency (Experiment 1)

More germane to our hypothesis, there was also evidence for an interaction between Verb type and Sentence voice in the critical region (95% CI: [−0.21, −0.02]; $p_{MCMC} = 0.01$). To examine this interaction, we considered two kinds of nested models, one which nested Verb type within Sentence voice (table 64), and one which nested Sentence voice within Verb type (table 65).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Verb type (active)	0.02	0.03	−0.03	0.08	0.21
Verb type (passive)	−0.09	0.03	−0.16	−0.03	< 0.00

Table 64: Verb type nested in Sentence voice model for critical region without frequency (Experiment 1)

For passive sentences, there was evidence for an effect of Verb type (95% CI: [−0.16, −0.03]; $p_{MCMC} < 0.00$), but there was no evidence for a similar effect in active sentences. The effect in passive sentences reflected that in passives, the subject experiencer condition took longer to read than the object experiencer condition.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Sentence voice (subjexp)	0.78	0.03	0.72	0.85	0
Sentence voice (objexp)	0.67	0.03	0.61	0.73	0

Table 65: Sentence voice nested in Verb type model for critical region without frequency (Experiment 1)

For both subject experiencer and object experiencer verbs, there was evidence that passives took longer to read than actives; this is again certainly because of the additional word in the passive condition, so we do not consider it further.

Overall, the models for the critical region revealed that Verb type interacted with Sentence voice. This was because in passive sentences, subject experiencer verbs produced longer region RTs than object experiencer verbs, while there was no reliable difference in active sentences.

Appendix E.1.2 Spillover region

In the spillover region, there was evidence for a main effect of Sentence voice (95% CI: [−0.13, −0.03]; $p_{MCMC} < 0.00$), as well as for an interaction between Verb type and Sentence voice (95% CI: [−0.34, −0.16]; $p_{MCMC} = 0.00$). The main effect of Sentence voice reflected the fact that in the spillover region, passives were read more quickly on average than actives. However, in the region based analysis, it is difficult to directly interpret any main effect of Sentence voice, as the different voice conditions had different numbers of words. For this reason, we refrain from interpreting this effect. To investigate the interaction, we again used a model with Verb type nested in Sentence voice (table 67) as well as a model with Sentence voice nested in Verb type (table 68).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.79	0.07	6.64	6.94	–
Verb type	0.02	0.02	−0.02	0.06	0.19
Sentence voice	−0.08	0.02	−0.13	−0.03	< 0.00
Verb type × Sentence voice	−0.25	0.05	−0.34	−0.16	0.00

Table 66: Crossed model for spillover region without frequency (Experiment 1)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Verb type (active)	0.14	0.03	0.08	0.20	0.00
Verb type (passive)	−0.11	0.03	−0.17	−0.05	< 0.00

Table 67: Verb type nested in Sentence voice model for spillover region without frequency (Experiment 1)

The model with Verb type nested in Sentence voice revealed evidence for an effect of Verb type in both voices. However, crucially, these effects are in the opposite direction. This reflects the fact that in actives, object experiencer verbs produced longer spillover region RTs than subject experiencer verbs; while in passives, the opposite pattern occurred.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Sentence voice (subjexp)	0.05	0.03	−0.02	0.11	0.07
Sentence voice (objexp)	−0.20	0.03	−0.27	−0.13	0.00

Table 68: Sentence voice nested in Verb type model for spillover region without frequency (Experiment 1)

The model with Sentence voice nested in Verb type found evidence for a single effect of Sentence voice, with object experiencer verbs. This reflected the fact that participants' spillover region RTs for object experiencer verbs were longer in active sentences than in passive sentences.

The models regressing spillover region RTs overall revealed that Sentence voice interacted with Verb type. This was because of two reasons. First, object experiencer verbs led to greater difficulty in active sentences than subject experiencer verbs, and this pattern was reversed in passive sentences. Second, active sentences with object experiencer verbs produced longer spillover region RTs than did passive sentences with object experiencer verbs, while no reliable effect was found for subject experiencer verbs.

Appendix E.2 Voice frequency models

Figure 12 shows the regression of mean log region RTs on frequency.

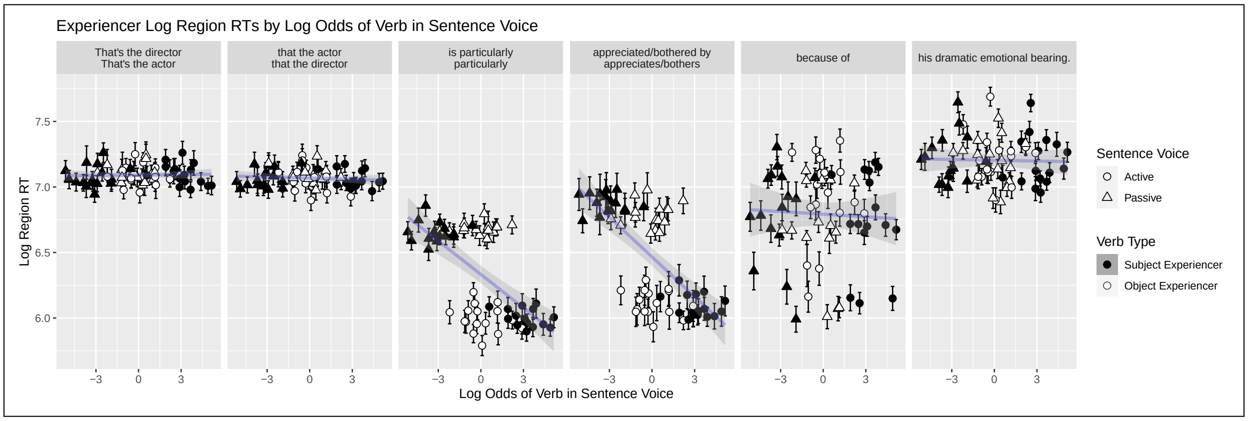


Figure 12: Regression of mean log region RTs on frequency (Experiment 1)

Appendix E.2.1 Critical region

At the critical region, there was evidence for an effect of frequency (95% CI: [−0.13, −0.09]; $p_{MCMC} = 0$). This reflected the fact that participants read this region on average more quickly when a verb was more likely to occur in the voice it occurred in than when it was less likely. That is, they read verbs that were more likely to occur in the active voice faster in the active voice than in the passive voice, and they read verbs that were more likely to occur in the passive voice faster in the passive voice than in the active voice. However, as figure 12 clearly shows, this effect is confounded with the fact that the critical region was read faster on average for every verb in the active voice than in the passive voice. This is because the critical region contained two words in the passive voice sentences, and only one word in the active voice sentences. For this reason, we believe this effect is not likely to be informative considered on its own.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.47	0.05	6.37	6.57	–
Sentence voice log odds	−0.11	0.01	−0.13	−0.09	0.00

Table 69: Frequency model for critical region (Experiment 1)

Appendix E.2.2 Spillover region

There was no evidence of an effect of frequency in the spillover region (95% CI: [−0.02, 0.00]; $p_{MCMC} = 0.08$).

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.79	0.08	6.64	6.94	–
Sentence voice log odds	−0.01	0.00	−0.02	0.00	0.08

Table 70: Frequency model for spillover region (Experiment 1)

Appendix E.3 Voice, Verb type, and Voice frequency models

Appendix E.3.1 Critical region

For the critical region, the model combining the interaction of the categorical predictors plus frequency revealed a main effect of Sentence voice (95% CI: [0.63, 0.77]; $p_{MCMC} = 0$). This was the same main effect that showed up in the model that just contained the categorical predictors, and was due to participants taking longer to read this region in passive sentences than in active sentences, certainly because passive sentences contained an extra word and not because of anything germane to our questions of interest.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.47	0.05	6.36	6.57	–
Verb type	−0.03	0.02	−0.08	0.01	> 0.05
Sentence voice	0.70	0.04	0.63	0.77	0.00
Sentence voice log odds	−0.01	0.01	−0.03	0.01	0.18
Verb type × Sentence voice	−0.07	0.07	−0.21	0.07	0.18

Table 71: Crossed model for critical region with frequency (Experiment 1)

Otherwise, there was no evidence for any other main effect or interaction. This result contrasted with the results of the model fit with only the categorical predictors, which additionally found evidence for an interaction between Verb type and Sentence voice; as well as with the frequency-only model, which found evidence for a main effect of frequency.

Appendix E.3.2 Spillover region

The model combining the categorical predictors and frequency for the spillover region revealed evidence for a main effect of Sentence voice (95% CI: [−0.16, −0.01]; $p_{MCMC} = 0.02$), as well as for an interaction between Verb type and Sentence voice (95% CI: [−0.38, −0.09]; $p_{MCMC} < 0.00$). This was consistent with results of the model fit using only categorical predictors, which revealed the same main effects and interaction, as well as with the frequency only model, which similarly found no effect of frequency in the spillover region. The main effect reflected that overall, participants took less time to read the spillover region in passives than in actives.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	6.79	0.08	6.64	6.95	–
Verb type	0.02	0.02	−0.02	0.06	0.19
Sentence voice	−0.09	0.04	−0.16	−0.01	0.02
Sentence voice log odds	−0.00	0.01	−0.02	0.02	0.42
Verb type × Sentence voice	−0.23	0.07	−0.38	−0.09	< 0.00

Table 72: Crossed model for spillover region with frequency (Experiment 1)

To examine the interaction, we fit models that resolved it by both Verb type and Sentence voice that also included frequency as a predictor. We provide tables here summarizing the results. These results revealed the same pattern of effects in the same directions as the nested models fit using only categorical predictors (in section 2.4.1), so we do not discuss them in further detail.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Verb type (active)	0.13	0.04	0.05	0.21	< 0.00
Verb type (passive)	−0.10	0.04	−0.18	−0.02	0.01

Table 73: Verb type nested in Sentence voice model for spillover region with frequency (Experiment 1)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Sentence voice (subjexp)	0.02	0.06	−0.11	0.15	0.38
Sentence voice (objexp)	−0.20	0.03	−0.27	−0.14	0.00

Table 74: Sentence voice nested in Verb type model for spillover region with frequency (Experiment 1)

Appendix E.4 Summary of effects

Type of predictors	Measure	Effects found	Explanation		
Categorical	Critical region RTs	Sentence voice	passive	>	active
		Verb type × Sentence voice	subjexp passive	>	objexp passive
	Spillover region RTs	Sentence voice	active	>	passive
		Verb type × Sentence voice	objexp active	>	subjexp active
			subjexp passive	>	objexp passive
			objexp active	>	objexp passive
Frequency	Critical region RTs	Sentence voice frequency	higher frequency	↔	lower RT
Both	Critical region RTs	Sentence voice	passive	>	active
	Spillover region RTs	Sentence voice	active	>	passive
		Verb type × Sentence voice	objexp active	>	subjexp active
			subjexp passive	>	objexp passive
			objexp active	>	objexp passive

Table 75: Summary of model results (Experiment 1)

Appendix E.5 Model comparisons

Appendix E.5.1 Critical region

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV × VT	0.0	0.0	0.0	0.0	78; 0
SV × VT + SFq	−1.9	1.3	−2.0	1.3	37; 0
SFq	−298.8	26.7	−298.4	26.8	80; 0

Table 76: Critical region RTs predictive crossed model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV + VT ∈ A + VT ∈ P	0.0	0.0	0.0	0.0	78; 0
SV + VT ∈ A + VT ∈ P + SFq	−1.6	1.4	−1.6	1.4	37; 0
SFq	−300.0	26.9	−299.6	26.9	81; 0

Table 77: Critical region RTs predictive Verb type nested in Sentence voice model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0.0	0.0	81; 0
VT + SV ∈ SE + SV ∈ OE + SFq	−2.5	1.3	−2.6	1.3	37; 0
SFq	−299.0	27.0	−298.6	27.0	82; 0

Table 78: Critical region RTs predictive Sentence voice nested in Verb type model comparisons (Experiment 1)

Predictive comparisons Looking at the predictive model comparisons for the region-based analyses for experiment 1, we can see that numerically, the models without frequency perform best, followed closely by the models that use both grammatical predictors and frequency. Since the absolute value of the two difference elpd_diff scores are less than 4, these models both perform roughly equally well at predicting the data. However, the model that uses only Sentence voice frequency performs much worse than the other two, since its elpd_diff scores are > 4, and they are large relative to the standard error of those differences.

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV × VT	0.0	0.0	0.0	0.0	78; 0
SV × VT + *SFq*	0.0	0.9	0.0	0.9	78; 0
SV × *VT* + SFq	0.0	0.0	0.0	0.0	37; 0
SFq	−299.0	26.2	−298.9	26.2	54; 0
SV + VT ∈ A + VT ∈ P	0.0	0.0	0.0	0.0	78; 0
SV + VT ∈ A + VT ∈ P + *SFq*	−0.7	0.9	−0.8	0.9	78; 0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	0.0	0.0	37; 0
SFq	−298.7	26.2	−298.7	26.2	53; 0
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0.0	0.0	81; 0
VT + SV ∈ SE + SV ∈ OE + *SFq*	−0.5	1.0	−0.5	1.0	79; 0
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	0.0	0.0	37; 0
SFq	−298.4	26.2	−298.4	26.2	55; 0

Table 79: Critical region RTs additive model comparisons (Experiment 1)

Additive comparisons Comparing the additive models, which compared models by adding a single set of fixed effects at a time without modifying the random effects structure, we see a consistent pattern: adding frequency as a predictor to the models using categorical predictors does not lead to any improvement in predictive utility, while in contrast, adding the categorical predictors to the frequency only models leads to a large improvement in predictive utility. However, this is certainly because of the fact that the critical region contains two words in the passive voice and one in the active voice. Thus, having direct access to the Sentence voice associated with a measure in this region is exceptionally useful in a way that is irrelevant to our research question.

Appendix E.5.2 Spillover region

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV × VT	0.0	0.0	0.0	0.0	83; 2
SV × VT + SFq	−1.2	1.1	−1.2	1.1	55; 0
SFq	−34.6	10.2	−33.8	10.3	88; 0

Table 80: Spillover region RTs predictive crossed model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV + VT ∈ A + VT ∈ P	0.0	0.0	0.0	0.0	86; 0
SV + VT ∈ A + VT ∈ P + SFq	−0.4	1.5	−0.5	1.5	55; 1
SFq	−31.5	10.1	−30.9	10.1	88; 0

Table 81: Spillover region RTs predictive Verb type nested in Sentence voice model comparisons (Experiment 1)

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0.0	0.0	81; 1
VT + SV ∈ SE + SV ∈ OE + SFq	−1.7	1.3	−1.4	1.3	55; 0
SFq	−36.1	10.5	−35.2	10.5	85; 0

Table 82: Spillover region RTs predictive Sentence voice nested in Verb type model comparisons (Experiment 1)

Predictive comparisons In general, across the predictive comparisons in the spillover region, we found the same pattern as in the critical region comparisons: models with the categorical predictors were similar to the models combining those and frequency, while the models with frequency had worse predictive utility (> 3 standard errors away from the best model).

Model	WAIC		LOO		*n* prob. obs.
	elpd_diff	se_diff	elpd_diff	se_diff	
SV × VT	0.0	0.0	0.0	0.0	83; 2
SV × VT + *SFq*	−0.5	0.4	−0.3	0.3	82; 1
SV × *VT* + SFq	0.0	0.0	0.0	0.0	55; 0
SFq	−28.4	7.9	−28.3	8.0	59; 0
SV + VT ∈ A + VT ∈ P	0.0	0.0	0.0	0.0	86; 0
SV + VT ∈ A + VT ∈ P + *SFq*	−0.3	0.3	−0.4	0.4	87; 0
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	0.0	0.0	55; 0
SFq	−28.4	7.9	−28.3	7.9	59; 0
VT + SV ∈ SE + SV ∈ OE	0.0	0.0	0.0	0.0	81; 1
VT + SV ∈ SE + SV ∈ OE + *SFq*	−0.7	0.5	−0.5	0.5	80; 1
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	0.0	0.0	55; 0
SFq	−28.6	7.9	−28.5	8.0	60; 0

Table 83: Spillover region RTs additive model comparisons (Experiment 1)

Additive comparisons We also found as similar pattern as in the critical region for the spillover region additive comparisons: models that added frequency did not suffer for it, but adding the categorical predictors to the frequency only model greatly improved performance (> 3 standard errors from the better model).

Appendix F. 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 (and also excluding regions where any individual word met either of these criteria). These led to a similar pattern of results as what we found in the text. We summarize the results in table 84. In case we strike out a main effect but not the corresponding explanation, it means that there was still evidence for the pairwise effect in the nested model. (And of course, filtering reading times did not impact the Question accuracy models, which were not based on reading time data.) A main effect that we found evidence for in the models fit using filtered RTs, which we did not find evidence for in the models using unfiltered RTs, is bolded.

We note the following differences: evidence for an interaction between Verb type and Sentence voice in the critical region goes away in the model without frequency (though there is still evidence for the subjexp passive > objexp passive effect in the nested model). There was evidence for an additional main effect of Sentence voice found in the model without frequency for the spillover word RTs. There was no evidence of an effect of Sentence voice for subject experiencer verbs in the spillover word RT model without frequency. There was no evidence for an effect of frequency in the frequency only model for the spillover word RTs. There was no evidence for an effect of Verb type for passive sentences in the model using both kinds of predictors for the spillover region RTs. Finally,

Type of predictors	Measure	Effects found	Explanation		
Categorical	Critical region RTs	Sentence voice	passive	>	active
		~~Verb type × Sentence voice~~	subjexp passive	>	objexp passive
	Spillover region RTs	Sentence voice	active	>	passive
		Verb type × Sentence voice	objexp active	>	subjexp active
			subjexp passive	>	objexp passive
			objexp active	>	objexp passive
	Critical verb RTs	Sentence voice	active	>	passive
	Spillover word RTs	**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	Critical region RTs	Sentence voice frequency	higher frequency	↔	lower RT
	Spillover word RTs	~~Sentence voice frequency~~	~~higher frequency~~	~~↔~~	~~lower RT~~
	Question accuracy	Sentence voice frequency	higher frequency	↔	higher accuracy
		Question voice frequency	higher frequency	↔	lower accuracy
Both	Critical region RTs	Sentence voice	passive	>	active
	Spillover region RTs	Sentence voice	active	>	passive
		Verb type × Sentence voice	objexp active	>	subjexp active
			~~subjexp passive~~	~~>~~	~~objexp passive~~
			objexp active	>	objexp passive
	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

Table 84: Summary of model results with filtered RTs (Experiment 1)

the marginal evidence for an effect of Verb type for passive sentences in the model using both kinds of predictors for the spillover word was no longer present. Though this does weaken the pattern of results somewhat (in particular, the only evidence for the full crossover interaction when filtering RTs only comes from Question accuracy, which of course did not undergo filtering), the overall pattern is not substantially changed. We also note that evidence for one of the effects of frequency was similarly eliminated by this filtering process, so it did not affect only the grammatical predictors. The model comparisons revealed similar patterns to the comparisons reported in the text, with the exception that the comparisons for the spillover word no longer distinguished the predictive capabilities of the frequency-only models from the models that used categorical predictors or both kinds of predictors (< 2 standard errors away from the best model in all cases). Data for these comparisons can be found in our supplementary material.

However, we believe there is good reason to be skeptical of over-interpreting the differences that filtering out some of the RTs caused, particularly when considering the pattern of exactly which RTs were filtered out. Table 85 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 region/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.

Region/Word	Verb type	Sentence voice	
		Active	Passive
Critical region	SubjExp	599 (3; 6)	565 (8; 35)
	ObjExp	588 (4; 16)	597 (1; 10)
Spillover region	SubjExp	596 (6; 6)	578 (5; 25)
	ObjExp	574 (6; 28)	602 (3; 6)
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 85: Observations per condition after filtering RTs < 100 ms or > 3000 ms (Experiment 1)

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 across three out of four measures; the single exception is that there are not many subject experiencer passive exclusions for the critical verb (which did not provide the strongest evidence for our arguments, at any rate). These are precisely the conditions that participants showed the most difficulty with in comprehension questions. As such, it is unsurprising that RTs for the verb and post-verbal material in these conditions would be longer, if participants are 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.

Appendix G. Experiment 2: Voice, Verb type, and Voice frequency model

As in experiment 1, for experiment 2 the models fit with only categorical predictors and the model using only frequency both found evidence for some effects of those predictors. Thus, we fit models combining both kinds of effects to determine the potential relevance of both sources of variance, as well as to set us up for a comparison of models.

Table 86 reports the results of the model crossing our categorical predictors and adding frequency.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Intercept	2.07	0.20	1.69	2.46	–
Verb type	−0.40	0.19	−0.79	−0.03	0.02
Sentence voice	0.43	0.27	−0.12	0.95	0.06
Sentence voice log odds	−0.08	0.07	−0.22	0.05	0.11
Verb type × Sentence voice	−1.86	0.40	−2.65	−1.08	0.00

Table 86: Crossed model for Response with frequency (Experiment 2)

Like the model using only categorical predictors, the model combining the categorical predictors with frequency found evidence for a main effect of Verb type (95% CI: [−0.79, −0.03]; $p_{MCMC} = 0.02$), and evidence for an interaction of Verb type and Sentence voice (95% CI: [−2.65, −1.08]; $p_{MCMC} = 0$). Unlike this model, there was also marginal evidence for an effect of Sentence voice (95% CI: [−0.12, 0.95]; $p_{MCMC} = 0.06$), which seems to reflect a trend for overall more "grammatical" judgments in response to active sentences than to passive sentences (active: $\mu = 0.818, \sigma = 0.0121$; passive: $\mu = 0.804, \sigma = 0.0125$). However, these effects were qualified by the interaction between Verb type and Sentence voice, which we examine below. Unlike the frequency only model, there was no evidence found for an effect of frequency (95% CI: [−0.22, 0.05]; $p_{MCMC} = 0.11$).

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 found evidence for all four effects. Model results are reported in tables 87 and 88.

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Verb type (active)	−1.79	0.37	−2.53	−1.09	0.00
Verb type (passive)	0.95	0.30	0.35	1.54	< 0.01

Table 87: Verb type nested in Sentence voice model for Response with frequency (Experiment 2)

Covariate	Estimate	Est. Error	l-95% CI	u-95% CI	p_{MCMC}
Sentence voice (subjexp)	1.80	0.46	0.91	2.71	0.00
Sentence voice (objexp)	−0.83	0.28	−1.40	−0.28	< 0.00

Table 88: Sentence voice nested in Verb type model for Response with frequency (Experiment 2)

Appendix H. Experiment 2: Results of WAIC comparisons

This appendix presents the results of the WAIC comparisons for experiment 2 (see Appendix C and section 3.2.4). As with experiment 1, the results of the LOO comparisons (presented in section 3.2.4) and the results of the WAIC comparisons were similar; we simply present the tables here with no further commentary.

Appendix H.1 Predictive comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT + SFq	0.0	0.0	60
SV × VT	−0.8	1.7	18
SFq	−30.9	7.6	67

Table 89: Response predictive crossed model WAIC comparisons (Experiment 2)

Model	elpd_diff	se_diff	n prob. obs
$SV + VT \in A + VT \in P + SFq$	0.0	0.0	67
$SV + VT \in A + VT \in P$	−2.1	2.1	18
SFq	−33.1	8.0	71

Table 90: Response predictive Verb type nested in Sentence voice model WAIC comparisons (Experiment 2)

Model	elpd_diff	se_diff	n prob. obs
VT + SV ∈ SE + SV ∈ OE + SFq	0.0	0.0	67
VT + SV ∈ SE + SV ∈ OE	−1.8	2.0	18
SFq	−32.0	8.2	71

Table 91: Response predictive Sentence voice nested in Verb type model WAIC comparisons (Experiment 2)

Appendix H.2 Additive comparisons

Model	elpd_diff	se_diff	n prob. obs
SV × VT + *SFq*	0.0	0.0	60
SV × VT	−1.2	1.2	60
SV × *VT* + SFq	0.0	0.0	18
SFq	−23.4	6.3	19
SV + VT ∈ A + VT ∈ P + *SFq*	0.0	0.0	67
SV + VT ∈ A + VT ∈ P	−2.5	1.6	62
SV + *VT* ∈ *A* + *VT* ∈ *P* + SFq	0.0	0.0	18
SFq	−24.0	6.7	17
VT + SV ∈ SE + SV ∈ OE + *SFq*	0.0	0.0	67
VT + SV ∈ SE + SV ∈ OE	−1.8	1.5	64
VT + *SV* ∈ *SE* + *SV* ∈ *OE* + SFq	0.0	0.0	18
SFq	−24.0	6.7	19

Table 92: Response additive model WAIC comparisons (Experiment 2)

Appendix I. 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 available 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 were defined as they were for the region-based analysis of SPR RTs presented in Appendix E.

We have not run any statistical tests on 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 13 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.

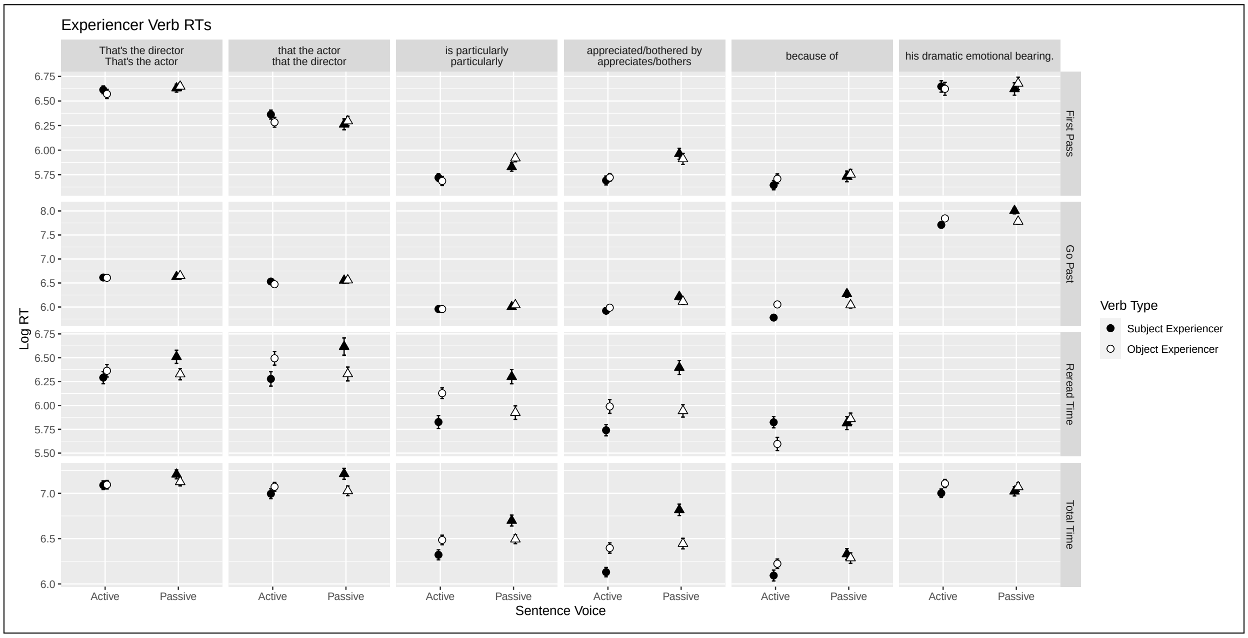


Figure 13: Log RTs for four measures by region, Verb type, and Sentence voice (Eye-tracking Pilot)

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 14 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.

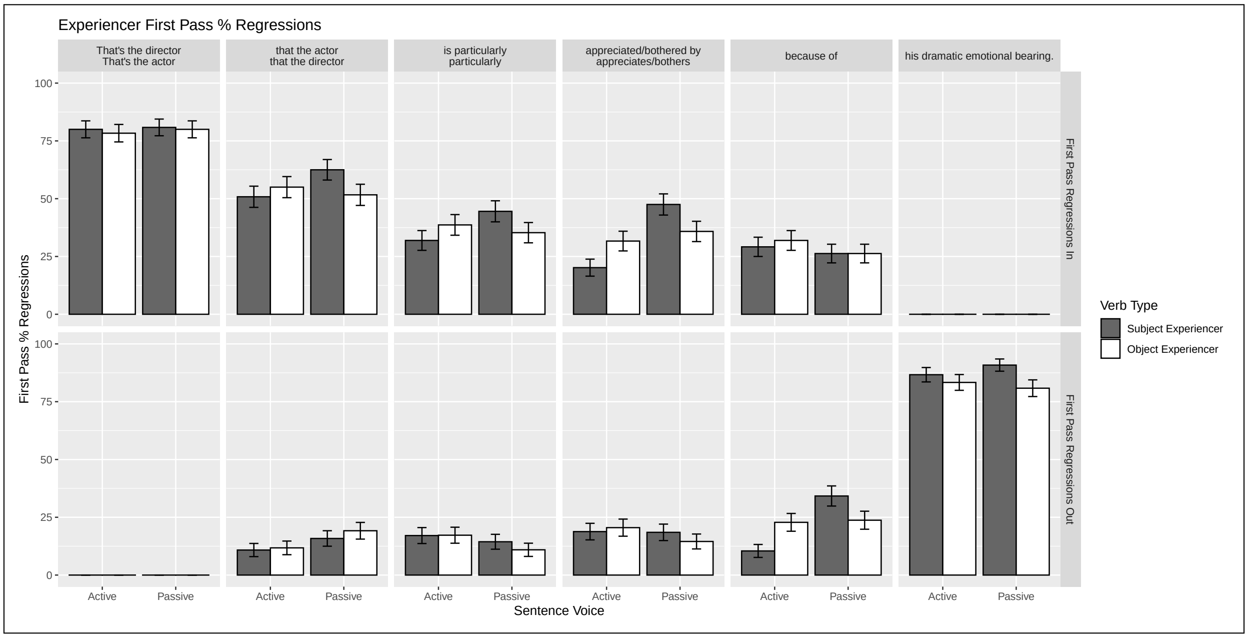


Figure 14: % First pass regressions in/out by region, Verb type, and Sentence voice (Eye-tracking Pilot)

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.

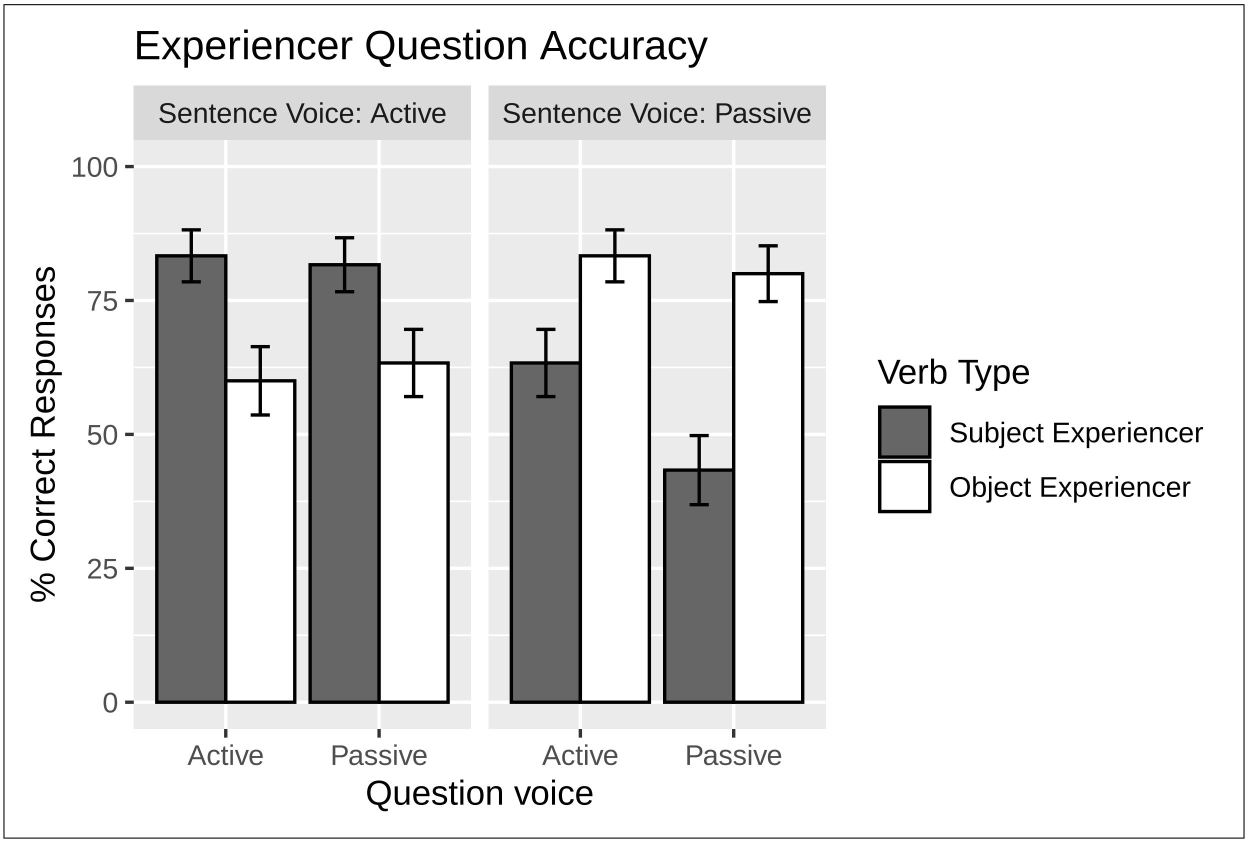


Figure 15: Question accuracy by Verb type, Sentence voice, and Question voice (Eye-tracking pilot)

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.