How to Sneeze a Napkin off the Table: Understanding Grammatically Creative, Coerced Sentences in Real Time

Tobias Ungerer (University of Toronto Scarborough & Concordia University), Caitlyn Antal (McGill University), and Roberto G. de Almeida (Concordia University)

Abstract: While numerous studies have examined how speakers understand newly coined words and novel figurative expressions, it remains largely unknown how grammatically creative sentences are processed in real time. In two reading experiments, we investigated how speakers comprehend instances of valency coercion, where a verb combines with noncanonical grammatical arguments (e.g., *Frank sneezed his napkin off the table*). Experiment 1 (N = 80), which included a preregistered replication (N = 120), employed the "maze" variant of the self-paced reading task. We found that coerced sentences, compared with prototypical (uncreative) controls, produced immediate processing difficulty after the verb, which was, however, rapidly alleviated at the prepositional phrase. Experiment 2 (N = 55), using eye-movement recordings, showed that the processing difficulty in coerced sentences was more successfully resolved than in fully anomalous controls, and that this resolution occurred both at temporally early and later stages of processing. Our results demonstrate that verb argument structure composition is flexible and computed during real-time incremental sentence comprehension. Comprehenders understand creative verb-argument combinations by rapidly integrating information from the verb and its clausal context.

Language is a hoard of human creativity. Not only do speakers frequently coin new words, such as *Brexit* or *metaverse*, but more numerous yet are the ways in which familiar words can be combined in novel and creative ways (Chomsky, 1966). Such combinatorial creativity can, on the one hand, be achieved on a purely semantic level, for instance by connecting seemingly unrelated words to express new figurative meanings, as in *her love is a shooting star*. On the other hand, it can also arise from a breach of structural principles, for example when a verb is combined with grammatical arguments with which it usually does not occur, as illustrated in (1) (adapted from Goldberg, 1995).[1]

(1) a. *Frank sneezed his napkin off the table.*

'Frank caused his napkin to fall off the table by sneezing on it.'

b. *They laughed the poor guy out of the room.*

'They caused the poor guy to leave the room by laughing at him.'

In these examples, prototypically intransitive verbs (*sneeze*, *laugh*) are combined with a direct object and a prepositional phrase to express an action that brings about a movement, as illustrated by the respective paraphrases. However, while these creative verb-argument linking patterns have been discussed from various theoretical perspectives (Audring & Booij, 2016; Boas, 2011; Goldberg, 1995; Lauwers & Willems, 2011; Michaelis, 2005; Müller & Wechsler,

[1] Chomsky's (1966) notion of "creative aspect of language use" embodies two notions of linguistic creativity. One is that linguistic expressions are free from the control of stimuli; the second is that the linguistic system (viz., the grammar) is productive in the sense that "language provides finite means but infinite possibilities of expression constrained only by rules of concept formation and sentence formation (…)" (Chomsky, 1966, p. 29). This latter notion is closer to the view of grammatical creativity that we employ in the present paper, even though we further extend it to non-canonical, and thus potentially rule-breaking, verb-argument combinations as exemplified in (1) (see also Bergs, 2019).

2014), the question of how language users process them in real time has been scarcely broached (Busso et al., 2021). As a result, fundamental aspects of the interplay between canonical verb representations and their use in non-canonical sentence contexts remain unaddressed, such as: How are grammatically creative sentences processed during rapid "on-line" (i.e., real-time) comprehension? How quickly are comprehenders able to resolve the conflict between properties of the verb and the clausal construction? How does processing of these creative structures differ from canonical, structurally unmarked sentences and from fully anomalous, uninterpretable ones?

In the present paper, we report on two experiments investigating the comprehension of grammatically creative sentences, such as the ones in (1). In Experiment 1A and its preregistered replication in Experiment 1B, we used the "maze" variant of self-paced reading to study the time-course of processing on a word-by-word basis. In Experiment 2, we relied on eye-movement recordings to investigate the mechanisms of reading comprehension employing a more naturalistic technique. By providing converging evidence from these paradigms, our experiments address a key type of grammatical creativity—when sentence structure appears to conflict with canonical or default verb-argument structure.

Grammatical Creativity: An Understudied Phenomenon

The study of creative ideas, usually defined as ideas that are both novel and effective (Runco & Jaeger, 2012), has a long history in the cognitive sciences (Jones, 2015; Kaufman & Sternberg, 2019). Language not only serves as a fundamental medium for expressing creative thoughts, but the linguistic system itself is subject to frequent innovative change. Language users introduce novelty at all levels of the linguistic system, including lexical-morphological creativity (e.g., novel words, morphological blends, or compounds; Munat, 2015), semantic and pragmatic

creativity (e.g., novel metaphors, irony, or slang; Hidalgo-Downing, 2015), and grammatical creativity (Vogel, 2023), which includes modifications of canonical phrase or sentence structure as in (1) above.

From a descriptive perspective, lexical, semantic, pragmatic, and grammatical creativity have been well documented since early stylistic studies of creative, and especially poetic, language (e.g., Greenfield, 1967; Thorne, 1965). In contrast, there is a clear asymmetry in the extent to which the different types of creativity have been investigated in psycholinguistics. Much of the experimental literature on creative language processing has focused on semantic or pragmatic creativity in figurative language, and especially on novel metaphors (for a review, see Holyoak & Stamenković, 2018). The way in which the processing of these expressions differs from conventional metaphors has been studied with diverse methods, including cross-modal priming (e.g., Blasko & Connine, 1993), preference ratings (Bowdle & Gentner, 2005), self-paced reading (Horvat et al., 2022), eye-tracking (Ashby et al., 2018), event-related brain potentials (Arzouan et al., 2007), and functional-magnetic resonance imaging (Cardillo et al., 2012). With respect to lexical-morphological creativity, most studies have focused on the processing of novel compound words, which has been investigated using both behavioral (e.g., Coolen et al., 1993; Libben et al., 1999; Pollatsek et al., 2011) and electrophysiological methods (Bader et al., 2010; Meßmer et al., 2021).

In contrast, much less experimental work has addressed the processing of grammatical creativity. Most previous research on grammatical violations and atypical structures has focused on phenomena that, in our view, cannot be regarded as creative. There has been considerable work, especially in the electrophysiological literature (e.g., Kim & Gilley, 2013; Osterhout & Holcomb, 1992), on how ungrammatical sentences are processed, such as the omission of an

obligatory object argument in (2a). But these sentences are not creative given that the anomaly does not serve a functional purpose and is thus not "effective" (see the above definition of creativity). Other studies (Luka & Barsalou, 2005; Luka & Choi, 2012) have investigated "moderately grammatical" sentences such as (2b), described as "sentences that were grammatical, but would likely be revised by a good writer or editor" (Luka & Barsalou, 2005, p. 452). But such structures are neither genuinely novel nor particularly effective and thus do not illustrate deliberate creative use. Similar remarks apply to studies of non-canonical thematic role orderings, such as the unusual (but grammatical) occurrence of the recipient argument before the agent in (2c) (Manouilidou & Almeida, 2013; Rösler et al., 1998), or temporary syntactic ambiguities in garden-path sentences like (2d) (Christianson et al., 2017; Frazier & Rayner, 1982). Even though these structures may be infrequent and therefore difficult to process, they are nevertheless fully licensed by the grammar and do not fulfil the novelty criterion of creative language.

(2) a. *The woman persuaded to answer the door.*

b. *We hate to bake pies anymore.*

c. *Dann hat [dem Sohn]$_{recipient}$ [der Vater]$_{agent}$ [den Schnuller]$_{theme}$ gegeben.*
'(lit.) Then has to the son the father the pacifier given.'

d. *The horse raced past the barn fell.*

In contrast to these other research strands, our study examines the processing of a type of novel but interpretable (and thus "effective") grammatical structures, as we discuss next.

Research on Valency Coercion

Our study is concerned with cases such as (3), repeated from above, where a prototypically intransitive verb is combined with additional grammatical arguments, thus inviting a caused-motion interpretation in which Frank's sneezing caused the napkin to fall off the table (Goldberg, 1995).

(3) *Frank sneezed his napkin off the table.*

In the theoretical literature, this phenomenon has been discussed under varying labels, including accommodation (Goldberg, 1995), event composition (Levin & Rappaport Hovav, 2005), type shifting (De Swart, 1998; Michaelis, 2004), and coercion (Audring & Booij, 2016; Lauwers & Willems, 2011; Michaelis, 2005). Here, we adopt the coercion account, according to which the above example involves "a contextually licensed repair of a combinatorial conflict" (Lukassek et al., 2017, p. 805). Specifically, we use Busso et al.'s (2020, 2021) term "valency coercion" because the conflict arises as a verb is "forced" into an argument structure pattern that differs from its canonical argument-linking profile (or valency).

Valency coercion contrasts with other types of coercion that are more strongly driven by semantic rather than grammatical factors. The most widely studied of these are complement coercion, where, for instance, the canonical interpretation of a noun as an entity is shifted to an activity reading, as in *Mary began the novel* (suggesting an interpretation such as 'Mary began reading the novel'),[2] and aspectual coercion, where a temporally bounded activity may be

[2] But see de Almeida and Dwivedi (2008) for an account of this phenomenon in terms of semantic indeterminacy and pragmatic enrichment rather than semantic coercion.

coerced into a continuous reading, as in *The horse jumped until dawn* (with an interpretation that 'it jumped repeatedly'; De Swart, 1998; Pustejovsky, 1995). Other types of coercion include intensional transitives, such as *John wanted a beer* ('wanted to drink a beer'; Delogu et al., 2010), inchoative coercion, as in *Within 2 minutes, the boy was asleep* ('came to be asleep'; Brennan & Pylkkänen, 2010), and concealed questions, for example *The announcer guessed the winner of the contest* ('guessed who the winner was'; Harris et al., 2008). Compared with valency coercion, however, these phenomena appear to involve a lesser deviation from canonical rules and thus do not, in our view, constitute equally clear instances of creative language.

Whereas substantial research has investigated the real-time processing of these other types of coercion, and aspectual coercion in particular (e.g., Antal & de Almeida, 2021; Brennan & Pylkkänen, 2008; de Almeida et al., 2016; Kuperberg et al., 2010; Piñango et al., 2006), processing-related research on valency coercion is strikingly absent. Experimental work thus far has largely focused on investigating factors that influence the acceptability of valency coercion, relying on off-line methods (Busso et al., 2018, 2020; Perek & Hilpert, 2014; Yoon, 2016, 2019). For example, Busso et al. (2018, 2020) showed that the acceptability of coerced sentences in Italian depends on properties of the clause-level construction. If the construction typically combines with semantically similar verbs, it more liberally allows the coercion of new verbs, presumably because it is easier to classify the creative uses as instances of the existing construction. Perek and Hilpert (2014) tested how English and French second-language learners judge instances of valency coercion in German, demonstrating that the acceptability ratings varied depending on whether a corresponding grammatical structure exists in participants' first language or not. Yoon (2019) found that coerced sentences in Korean were judged as more acceptable when participants had previously been exposed to other coerced examples, suggesting

that comprehenders' "tolerance" towards valency coercion can, at least to some extent, be primed.

Busso et al.'s (2021) study constitutes, to our knowledge, the only attempt to examine valency coercion relying on a more on-line technique. In their study, participants first read Italian prime sentences such as (4) and then performed a lexical decision task on a target verb. Target verbs came in three types: construction associates, which were related to the intended meaning of the prime sentence (e.g., *to say* for (4), given that the sentence expresses an act of speech); lexical associates, which were related to the canonical sense of the prime verb, but which did not capture its coerced meaning (e.g., *to hum*); and unrelated verbs (e.g., *to age*).

(4) *Giovanni fischietta che verrà domani.*

'Giovanni whistles that he will arrive tomorrow.'

Busso et al. (2021) found that participants were faster at recognizing construction associates than both lexical associates and unrelated target verbs. This suggests that participants successfully decoded the coerced prime expressions, and that the novel meaning that prime verbs like *whistle* acquired in the coercing context was subsequently more accessible (as suggested by the priming effect) than the canonical meaning of these verbs.

Taken together, these studies demonstrate that comprehenders are able to interpret instances of valency coercion, and that their interpretation is affected both by properties of the clausal construction and by their own linguistic background. Crucially, however, the results do not elucidate the processes that underlie the on-line (i.e., real-time) processing of such grammatically creative sentences. Even Busso et al.'s (2021) findings only shed light on how

lexical verbs are processed *after* coerced sentences with related meanings, rather than how the coerced structures themselves are processed. Notice that participants in their study had 4000 milliseconds (ms) to read the sentence, with a 1000 ms fixation point before the target word appeared for lexical decision. Thus, while their study represents the first attempt to understand the process of comprehension of valency coercion, it does so only indirectly, by relying on faster responses to the construction-related verbs in contrast to other verb types. Moreover, the time taken between sentence reading and lexical decision does not capture the moment-by-moment processing of the verb and the integration—or even rejection—of its atypical arguments. As a result, the current state of the literature provides no evidence about the time-course of valency coercion processing, and specifically about the way in which comprehenders may resolve the difficulty posed by these grammatically creative sentences. The present study addresses this phenomenon by examining the real-time processing of valency coercion during reading. We next turn to how prior work on the processing of syntactic and semantic violations may inform our predictions.

Comprehension of Valency Coercion: Predictions and Open Questions

Our study focuses on instances of valency coercion in sentences with a postverbal noun phrase (NP) and locative prepositional phrase (PP), as in (5a) and (6a). We compare these *coerced* sentences with two control conditions: *prototypical* sentences, as in (5b) and (6b), which contain transitive verbs that canonically express caused motion; and *anomalous* sentences, as in (5c) and (6c), which feature intransitive verbs that are difficult to construe as a movement-inducing action.

(5) a. *Frank sneezed his napkin off the table.*

b. *Frank pushed his napkin off the table.*

c. *Frank arrived his napkin off the table.*

(6) a. *Sharon yelled her husband out of the kitchen.*

b. *Sharon shoved her husband out of the kitchen.*

c. *Sharon relied her husband out of the kitchen.*

To derive predictions for our experiments, we considered the incremental way in which comprehenders encounter the coerced stimuli, focusing on two sentence regions. First, the coerced sentences induce a temporary anomaly at the postverbal NP. They do this in two different ways: In (5a), the prototypically intransitive verb *sneeze* is combined with an unlicensed direct object (*his napkin*).[3] This is a violation of the verb's grammatical requirements, specifically its transitivity or subcategorization frame (Chomsky, 1965). In (6a), in contrast, the verb *yell*, which canonically only selects inanimate objects (e.g., *Sharon yelled her response*), is combined with an animate object (*her husband*). This animacy violation is a classic example of a violation of selectional restrictions and thus driven by semantic rather than grammatical cues (Katz & Fodor, 1963).

Previous experimental research provides strong evidence that both types of violations lead to processing difficulty. We focus here in particular on studies of reading comprehension that allow for close comparisons with our experiments. Evidence on transitivity violations comes from self-paced reading (Mitchell, 1987) and eye-tracking research (Staub, 2007; van Gompel &

[3] Unergative verbs like *sneeze* allow for cognate objects (e.g., *Frank sneezed a mighty sneeze*), but corpus data show that these transitive uses are exceedingly rare (see Materials and Norming). Moreover, an eye-tracking study of sentences with transitivity violations (Staub, 2007) found no processing difference between unergative verbs and unaccusative verbs, the latter of which do not allow cognate objects.

Pickering, 2001) that investigated sentences with unergative intransitive verbs, such as *struggled* in (7), which are comparable to the verbs in our coerced sentences.

(7) *When the dog struggled the vet and his new assistant took off the muzzle.*

Our concern here is not with the garden-path effect that occurs in (7), but merely with the fact that, across all studies, intransitive verbs led to slower reading at the postverbal NP (*the vet*), compared with a control condition of transitive verbs. In Staub's (2007) study, this processing difficulty emerged at a temporally early stage, when participants first read the NP, and it increased their likelihood to regress (i.e., look back) to earlier sentence regions.[4]

Violations of selectional restrictions, and of verb argument animacy in particular, have also been found to give rise to processing difficulty. Two eye-tracking studies (Warren et al., 2015; Warren & McConnell, 2007) provide evidence that in sentences like (8), in which the verb *blackmail* selects animate objects but is followed by an inanimate NP, reading times are increased at *spaghetti* and the following words (see also Paczynski & Kuperberg, 2012, for evidence from event-related brain potentials during reading). Similar to transitivity violations, the effects emerged both in temporally early and later eye-movement measures, suggesting that processing was immediately disrupted and led to increased re-reading efforts.

(8) *The man used a photo to blackmail the thin spaghetti yesterday evening.*

[4] As suggested by Staub (2007, Experiment 2), the difficulty at *the vet* may not arise from the transitivity violation, but because comprehenders (correctly) interpret the *vet* as the subject of the following main clause, with the difficulty stemming from the absence of a comma at the clause boundary. In our study, however, this interpretation is not available: Since our stimuli only contain a main clause, comprehenders cannot avoid the grammatical conflict between the intransitive verb and the following NP.

Based on these findings, we expected that the NP violations in our coerced sentences, compared with prototypical stimuli, would lead to processing difficulty at the NP in both experiments and, in our eye-tracking Experiment 2, would trigger increased re-reading efforts. We assumed that there could be differences between the two types of violations: In Experiment 1, which examined word-by-word reading, the effects of transitivity violations could already emerge at early stages of processing the NP—thus at the determiner *his* in (5a)—while animacy violations would only become apparent at the noun *husband* in (6a). In addition, the magnitude of processing difficulty could differ between violations, even though, to our knowledge, previous work has not explicitly compared the effect sizes of transitivity and animacy violations. Therefore, our main analyses in the experiments reported below compared coerced sentences (irrespective of violation type) with the control conditions, although we conducted additional analyses to investigate the effect of violation type. It is worth noting that our motivation for including both violation types was primarily practical, given the challenges of construing a sufficient number of experimental stimuli of only a single type. Beyond this, however, we also considered it a more robust test of valency coercion if it could be shown that the predicted effects would occur across sentences with different NP violations.

As for the comparison between coerced sentences and anomalous controls, we expected that both structures would give rise to difficulty at the NP. However, the effect could be stronger for anomalous sentences due to differences in real-word plausibility, which has been shown to affect reading difficulty (Staub et al., 2007; Warren & McConnell, 2007) in a graded way (Rayner et al., 2004). In particular, given that supportive context can alleviate the effects of implausibility (Filik, 2008; Warren et al., 2008), the preceding linguistic context in our stimuli (see Materials and Norming below) could allow participants to construe the unusual verb + NP

combinations in our coerced sentences as more plausible than the ones in the anomalous sentences, thus leading to greater difficulty in the latter cases.

The second key region of interest in our coerced stimuli is the clause-final PP, which specifies a locative goal (or destination). Critically, this is where the potential construal of the event as an instance of caused motion arises, which may allow comprehenders to resolve the preceding temporary anomaly (e.g., *sneeze the napkin*) and arrive at a plausible interpretation. Here, we outline one theoretical explanation of this resolution process, drawing on the usage-based accounts in which valency coercion has been most widely discussed (Boas, 2011; Busso et al., 2021; Goldberg, 1995; Perek & Hilpert, 2014; van Trijp, 2015). However, we will return to the question of how other theories of argument structure can account for the effects in the general discussion.

Goldberg (1995, pp. 53–55) bases her account of valency coercion on the framework of Construction Grammar, in which grammar is assumed to consist of constructional templates that combine elements of form and meaning (for an overview, see Ungerer & Hartmann, 2023). For example, the sentences in (5) and (6) all instantiate an abstract clausal caused-motion construction that links specific syntactic roles (subject, verb, object, "oblique" PP) to a schematic meaning ('X causes Y to move Z'). Instances of constructions are typically licensed if the participant roles encoded by their verb match the argument roles encoded by the construction. For instance, in (5b), the verb *push*, which lexically encodes three participant roles (a pusher, a "pushee," and a destination for pushing), matches the argument roles of the caused-motion construction (agent/cause, theme, and locative goal). In contrast, in the coerced example (5a), the verb *sneeze* only encodes a single participant role (a sneezer). In such special cases, Goldberg argues that the construction can contribute additional argument roles, thereby increasing the

valency of the verb and licensing its interpretation in terms of caused motion. This interpretation is felicitous if it can plausibly accommodate the lexical meaning of the verb (Goldberg, 1995, p. 159) given the specific discourse context (Boas, 2011). For instance, with sufficient contextual support, *sneeze* in (5a) may be plausibly construed as a manner of motion, whereas *arrive* in the anomalous example (5c) can hardly be construed in this way.[5]

Based on this theoretical account, we expect that the PP in our coerced stimuli will trigger a resolution process that alleviates their processing difficulty. In contrast, anomalous sentences should give rise to a persistent processing cost, given prior evidence that sentences with transitivity violations cause long-lasting disruptions that extend across the unexpected verbal arguments to the following sentence regions (Staub, 2007; van Gompel & Pickering, 2001). However, previous theoretical work provides no precise predictions about the time-course of resolution in coerced sentences: how quickly comprehenders overcome the initial combinatorial conflict, whether this occurs while they are processing the PP or via a retrospective reassessment after they have reached the end of the clause, and how exactly these processes manifest themselves in measures of reading time and eye movements. Our study addresses these open questions, along with the predictions outlined above, in two complementary reading settings, using the maze task and eye tracking.

Materials and Norming

Given the lack of processing studies on valency coercion in English, we created a new set of stimuli. We describe our full materials (along with norming procedures) here, before discussing the details of Experiment 1, which only used a subset of these materials.

---

[5] As a reviewer notes, caused-motion verbs such as *push* typically encode a less specific manner of motion than coerced verbs such as *sneeze*. This may trigger additional inference processes on the comprehenders' part (e.g., about how the pushing is done), which remain to be examined more closely in future work.

Stimulus Design

We created 24 items, consisting of text passages like the ones in (9) and (10). Each passage comprised two context sentences ((9a) and (10a)) and a third sentence containing the target expression ((9b) and (10b); the critical segments are boldfaced). The context sentences were intended to increase the plausibility of our coerced stimuli, building on prior evidence that contextual support facilitates the comprehension of creative expressions, such as novel metaphors (Bambini et al., 2016; Pynte et al., 1996). The targets consisted of a subject, a verb, an NP (which expressed a potential direct object), and a PP (which denoted a potential locative goal). Except for the verb, all content words of the targets already appeared in the context, ensuring that any effects at these regions would be due to structural rather than lexical factors. The targets were always embedded in a larger sentence, where they were preceded by a clause-initial adjunct and followed by a coordinated phrase starting with *and* plus a verb.

(9) a. *Frank swallowed a red chili pepper at the dinner table. Tears streamed from his eyes, and he reached blindly for his napkin.*

b. *Unable to control himself,* ***Frank (sneezed / pushed / arrived) his napkin off the table*** *and knocked over a few of the wine glasses.*

(10) a. *Sharon was arguing with her husband in the kitchen. They raised their voices as the discussion grew more and more heated.*

b. *In the end,* ***Sharon (yelled / shoved / relied) her husband out of the kitchen*** *and slammed the door with a loud bang.*

For each target sentence, we created three versions that differed in their verbs, resulting in 72 experimental items (see Appendix A for the full list of materials). "Coerced" sentences contained either a prototypically intransitive verb, such as *sneezed* in (9b), or a verb like *yelled* in (10b), which, if used transitively, canonically only occurs with inanimate objects (e.g., *yelled a greeting*) but was combined with an animate NP (*her husband*); see below for tests of these argument-linking properties. "Prototypical" control sentences contained (complex-)transitive verbs, such as *pushed* and *shoved*, which prototypically encode an action that brings about the motion of an object or person to a location. We selected verbs that did not lexically encode motion, but which could potentially be conceptualized as a movement-inducing action given the prior context. Finally, "anomalous" controls always featured intransitive verbs, such as *arrived* and *relied*, which in our view could not be construed as a movement-inducing action, even in the given context.

It is worth noting that our coerced verbs were unergative while our anomalous verbs comprised a mix of unaccusative and unergative verbs (e.g., *arrived* vs. *leaped*). However, eye-tracking evidence (Staub, 2007) suggests that this difference does not affect the processing of sentences containing transitivity violations. In addition, we examined whether the verbs differed in length (in letters) or lemma frequency (extracted from the *Corpus of American English*, COCA; Davies, 2008) across conditions. A linear regression analysis conducted in R (R Core Team, 2023) confirmed that there was no significant difference in length between prototypical, coerced, and anomalous verbs ($F(2, 71) = 1.12$, $p = .33$). Log-transformed frequency, however, differed significantly ($F(2, 71) = 7.49$, $p = .001$), with post-hoc comparisons indicating that prototypical verbs were more frequent than coerced verbs ($\beta = -1.60$, $SE = 0.42$, $t = -3.85$, $p < .001$) and marginally more frequent than anomalous verbs ($\beta = -0.95$, $SE = 0.42$, $t = -2.29$, $p =$

.06). To mitigate these effects, we included both length and frequency in our later statistical analyses at the verb. It should be noted, however, that our interest lay primarily in the postverbal regions, which were lexically identical across conditions.

In addition, we conducted two norming studies to ensure that our items had the desired grammatical features and were perceived in the intended way.

Argument-Linking Profiles of the Verbs

First, we investigated whether the use of verbs in our crucial coerced condition deviated from the verbs' prototypical argument-linking patterns and could thus be considered grammatically creative. For this purpose, we extracted a random sample of 100 instances per verb (in their past tense forms) from COCA (Davies, 2008), except for *tangoed*, which was only attested 14 times. We then annotated each instance for whether its verb was used intransitively or transitively and, in the case of the latter, whether the object phrase was animate or inanimate. Instances with verb particles (e.g., *shrugged off something/shrugged something off*), cognate objects (e.g., *smiled a friendly smile*), and reflexive objects (e.g., *sang oneself through it*) were counted as transitive. In contrast, instances with prepositional complements (e.g., *yelled at someone*), *that*-clauses (*yelled that…*), and direct speech in quotation marks (e.g., *yelled, "get out!"*) were counted as intransitive because these structures are syntactically quite distinct from the postverbal NPs used in our experiments. The full results are summarized in Appendix B.

Most verbs (20 out of 24) followed the expected pattern: Either the large majority of their corpus attestations were intransitive, but they were used transitively in our sentences (e.g., *sneezed*, *frowned*); or they were attested with inanimate objects in the corpus but occurred with animate objects in our materials (e.g., *yelled, sang*). Four verbs did not fully adhere to this pattern and were therefore more closely inspected. First, while *cheered* was most commonly

used intransitively in the corpus, 26% of its instances were combined with an animate object NP. Note, however, that most of these attestations involved the sense of 'make someone glad or happy,' while the sports context in our experimental item preactivated the distinct meaning of 'utter shouts of applause,' which was less frequently attested transitively in the corpus. Second, 7% of the corpus instances with *clapped* involved an animate NP; but again, note that these instances exclusively denoted a 'physical contact' scenario (*clapped someone on the back/shoulders*) that is distinct from the 'applaud' sense of the verb preactivated by the context of our experimental item. Third, *coughed* was attested with inanimate NPs in 25% of the corpus sample. But notably, these instances always combined with a particle (*coughed up/out*) and either denoted the substance expelled by coughing (e.g., *smoke*) or a metaphorical extension thereof (e.g., *coughed up cash*); while in our experiments, the object of coughing was an external object (*dust*), thus arguably deviating from the canonical use of the verb. Finally, 28% of corpus instances with *rumored* involved an animate NP. But all of them used the passive voice, thus suggesting that they instantiated an idiomatic construction (*someone is rumored to…*), which differs from the active transitive use of the verb in our experiments. Taken together, these findings suggest that, even though a few of our coerced items may potentially be less "deviant" than others, there were still semantic and/or syntactic cues in each case that made the sentences likely to be perceived as non-canonical by the time participants reached the noun of the NP.

In addition, we used our corpus results to classify verbs in terms of the type of NP violation they potentially give rise to. Specifically, we assumed that verbs that were used transitively in less than 5% of corpus instances may give rise to transitivity violations, while all other verbs (which had more transitive attestations, but mostly with inanimate NPs) may give rise to animacy violations. Even though the resulting set was not fully balanced (14 transitivity

violations vs. 10 animacy violations), we used the data for an exploratory analysis of violation type in our experiments.

Sentence Ratings

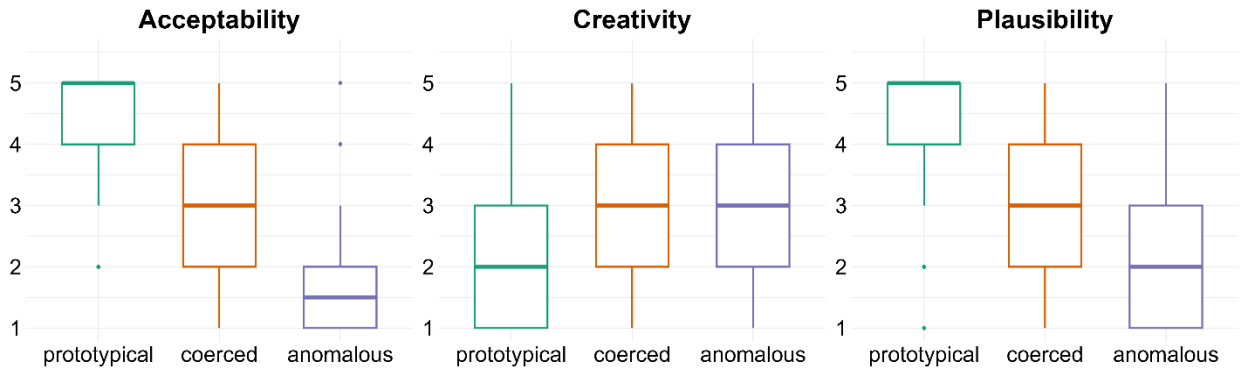
In a second step, we conducted a rating study in which 21 self-reported native speakers of English from the Concordia University community judged three global properties of our stimuli (embedded in their experimental contexts): acceptability (how (un)natural the sentence is), creativity (how ordinary or unusual it is), and plausibility (how (un)likely it is to be true in its context). Ratings were provided on a scale from 1 (lowest) to 5 (highest). We expected that coerced stimuli would receive intermediate acceptability ratings (because they are novel but interpretable), thus falling in between prototypical and anomalous sentences; that coerced sentences would be judged as more creative than prototypical ones; and that both coerced and prototypical sentences would be judged as highly plausible (because they align with the situational context).

Median rating scores for each item are included in Appendix A. Treating the ratings as ordinal data (Taylor et al., 2023), we analyzed them with cumulative link mixed models using the R package *ordinal* (Christensen, 2022). As illustrated in **Figure 1**, there were significant differences between sentence types in acceptability ($\chi^2(2) = 39.34$, $p < .001$), creativity ($\chi^2(2) = 11.18$, $p = .004$), and plausibility ($\chi^2(2) = 30.57$, $p < .001$). For acceptability, post-hoc comparisons indicated that prototypical sentences (Median = 5) were judged as more acceptable than both coerced sentences (Median = 3; $\beta = 3.96$, $SE = 0.64$, $z = 6.19$, $p < .001$) and anomalous sentences (Median = 1.5; $\beta = 6.79$, $SE = 0.80$, $z = 8.44$, $p < .001$), and that coerced sentences were judged as more acceptable than anomalous sentences ($\beta = 2.82$, $SE = 0.41$, $z = 6.87$, $p < .001$). In terms of creativity, prototypical sentences (Median = 2) were rated as less creative than

both coerced sentences (Median = 3; β = -1.99, *SE* = 0.55, z = -3.59, p = .001) and anomalous sentences (Median = 3; β = -2.31, *SE* = 0.72, z = -3.21, p = .004), while there was no difference between creative and coerced sentences (p = .78). Finally, plausibility ratings were similar to those for acceptability, with prototypical sentences (Median = 5) being rated as more plausible than both coerced sentences (Median = 3; β = 2.73, *SE* = 0.49, z = 5.55, p < .001) and anomalous sentences (Median = 2; β = 5.28, *SE* = 0.72, z = 7.34, p < .001), and coerced sentences being rated as more plausible than anomalous sentences (β = 2.55, *SE* = 0.42, z = 6.03, p < .001).

Figure 1

Norming Ratings for Acceptability, Creativity, and Plausibility by Sentence Type



Note. Horizontal lines are medians, boxes represent interquartile ranges, and whiskers extend to maximum/minimum points within 1.5x interquartile range.

The acceptability and creativity ratings confirm our predictions. At an item level, most stimuli followed this general trend, even though two coerced sentences (with *gestured* and *cheered*) were judged equally acceptable as their prototypical counterparts. We nevertheless decided to include the full stimulus set in our experiments and address potential item-level differences in the statistical analyses of our experiments. In contrast, plausibility ratings for

coerced sentences were lower than expected. This is potentially problematic because it leaves unclear whether differences between coerced and prototypical sentences in our experiments are due to real-world plausibility or structural factors. On the other hand, the novelty and structural anomaly of the coerced expressions may have indirectly affected their perceived plausibility. In addition, it is possible that participants did not clearly distinguish between acceptability and plausibility judgments, which would explain the very high correlation between the two ($r(70) = 0.93$, $p < .001$). We return to this issue in the discussion of our experiments.

Transparency and Openness

In the following sections, we report all data exclusions (if any), manipulations, and measures in the study, following the JARS guidelines (Appelbaum et al., 2018). The supplemental materials, including the data, code, model outputs, and plots for all analyses reported in this paper, are available at: https://doi.org/10.17605/OSF.IO/UQKP4 (Ungerer et al., 2025). Experiment 1 consisted of an exploratory study (Experiment 1A) and a preregistered replication (Experiment 1B); Experiment 2 was not preregistered.

Experiment 1A

In Experiment 1A, we investigated the word-by-word time course of valency coercion comprehension, using the maze task (Forster et al., 2009). In this task, participants read sentences word-by-word while choosing between a sensible sentence continuation and an incorrect distractor at every step. Compared with traditional self-paced reading, the maze task has been found to produce larger and statistically more robust processing effects (Boyce et al., 2020; Witzel et al., 2012). Moreover, since participants have to comprehend each word in order to choose the correct continuation, the effects are typically highly localized and rarely spill over to subsequent words (Boyce et al., 2020; Boyce & Levy, 2023).

Participants

Eighty participants were recruited online via the crowdsourcing platform Prolific (www.prolific.com). Participation was restricted to individuals who (a) reported English as their first and primary language, (b) currently lived in the UK, US, or Canada and had resided there for at least two years, (c) did not declare any language-related disorders or dyslexia, and (d) had an approval rating of 95% or higher. Participants were paid GBP 3 for the 15- to 20-minute experiment. All participants provided informed consent and were treated in accordance with the ethical standards adhered by Concordia University's Human Research Ethics committee (reference number 10000023).

Materials

The materials consisted of 24 text passages, each containing two context sentences followed by a target sentence (see (9) and (10) for examples). In this experiment, we only contrasted two versions of each target, a prototypical and a coerced expression, for a total of 48 critical items. We were not able to include the anomalous controls because the maze task requires the target words to be (at least marginally) sensible continuations and is thus not suited for investigating fully anomalous sentences.

For the purposes of the maze task, we created a distractor for each word of our target sentences, using Boyce et al.'s (2020) "Auto-maze." This tool draws on a large language model to select distractors that resemble the target words in length and frequency, but which are contextually inappropriate. Specifically, we created distractors for the prototypical sentences and then paired the same distractors with the corresponding coerced sentences. Distractors were, wherever possible, chosen such that they had high surprisal (at least 25 bits) and were more surprising than the correct word (by at least 5 bits). While this procedure does not guarantee that

the distractors are always impossible sentence continuations, the fact that participants were overall highly accurate in making the correct maze choices (see the results, below) suggests that the automatically generated distractors were sufficiently implausible.

For testing, we first distributed the stimuli over two lists to ensure that each participant only saw one version of the target sentences. We further split each list in half and tested participants on only one of the resulting four subsets, each containing 12 items (6 per condition).[6] This was to prevent participants from becoming overly familiar with the unusual structure of the coerced sentences. In addition to the experimental items, each list contained 24 filler items that had the same format as the experimental passages (consisting of two context sentences and one target) but with structures that differed from the experimental sentences.

Procedure

The experiment was run on the web platform PCIbex (Zehr & Schwarz, 2018). Participants completed two practice trials before starting the critical phase. In each trial, participants first read the two context sentences, presented together as a paragraph, and then pressed the space bar. On each subsequent screen, they saw a word of the target sentence displayed next to a distractor in the center of the screen, with the left-right position randomized. Participants decided which word was the correct sentence continuation by pressing "e" for the word on the left and "i" for the word on the right. When participants chose the incorrect option, an error message was displayed, after which they were allowed to correct their response. This allowed us to retain trials in which participants had made an error earlier on in the sentence (Boyce & Levy, 2023). Fifty percent of trials were followed by a comprehension question about

[6] Preliminary statistical analyses showed that results did not differ significantly between list halves. We therefore did not include the factor in the analyses reported below.

information that had been conveyed in the context sentences (but never the target sentence). Participants pressed “e” or “i” for yes and no, respectively, before receiving feedback on their response.

Data Preprocessing and Analysis

We first checked whether all participants were sufficiently accurate in the maze task (at least 80% correct maze choices, following Boyce & Levy, 2023) and the comprehension questions (at least 70% correct responses). Two participants were excluded as they failed to reach at least one of these thresholds; they were replaced with two new participants to obtain equal amounts of data across lists.

All analyses were conducted in R (R Core Team, 2023). For response times (RTs), we first excluded words for which participants had chosen the incorrect maze option (0.9% of the data). We then removed unnaturally short or long RTs below 100 ms and above 3000 ms (0.7% of the data). Finally, we excluded all datapoints that were more than 2.5 standard deviations away from participants’ individual means at a given word region (2.1% of the data). Since preliminary modeling showed that the RTs violated the assumption of normality of residuals, we log-transformed them to render their distribution more normal (Baayen & Milin, 2010).

We then analyzed the remaining RTs (7,051 datapoints) by fitting separate linear mixed regression models at each of the eight critical words, using the *lmerTest* package (Kuznetsova et al., 2017). The words ranged from the verb (e.g., *yelled*) to the first word following the PP (always *and*), which we treated as a potential spillover region. The main predictor variable of interest in all models was sentence type, sum-coded as prototypical (-0.5) vs. coerced (+0.5). We also added trial number (centered around the mean) as well as the interaction between trial number and sentence type because we expected that RTs could decrease over the course of the

experiment, and that this decrease could be stronger for the coerced sentences if participants became increasingly used to their unusual structure. For the models at the verb (the only word where the conditions differed lexically), we also added verb length (in number of letters) and the verbs' log-transformed frequency (derived from COCA; Davies, 2008). All models included maximal random effect structures (Barr et al., 2013) that led to model convergence. These included intercepts for subject and item in all cases, and random by-participant slopes for sentence type in all models except at the noun of the PP and at the spillover word. The *p*-values were computed with the Satterthwaite approximation for degrees of freedom (Kuznetsova et al., 2017). Apart from our main analysis, we also inspected RTs on a by-item level and conducted additional analyses to investigate the effects of NP violation type and our norming ratings. The details of each are explained in the "Results" section below.

For participants' accuracy in the maze task, we took the whole dataset into account (7,320 datapoints) and tried to fit logistic mixed models at each word that contained the same predictors as described above. Most of these models, however, either did not converge or produced singular fits. At the second preposition constituent of the PP and the spillover word, the models converged when random slopes were excluded. The full outputs of all statistical models are available in the supplemental materials.

Results

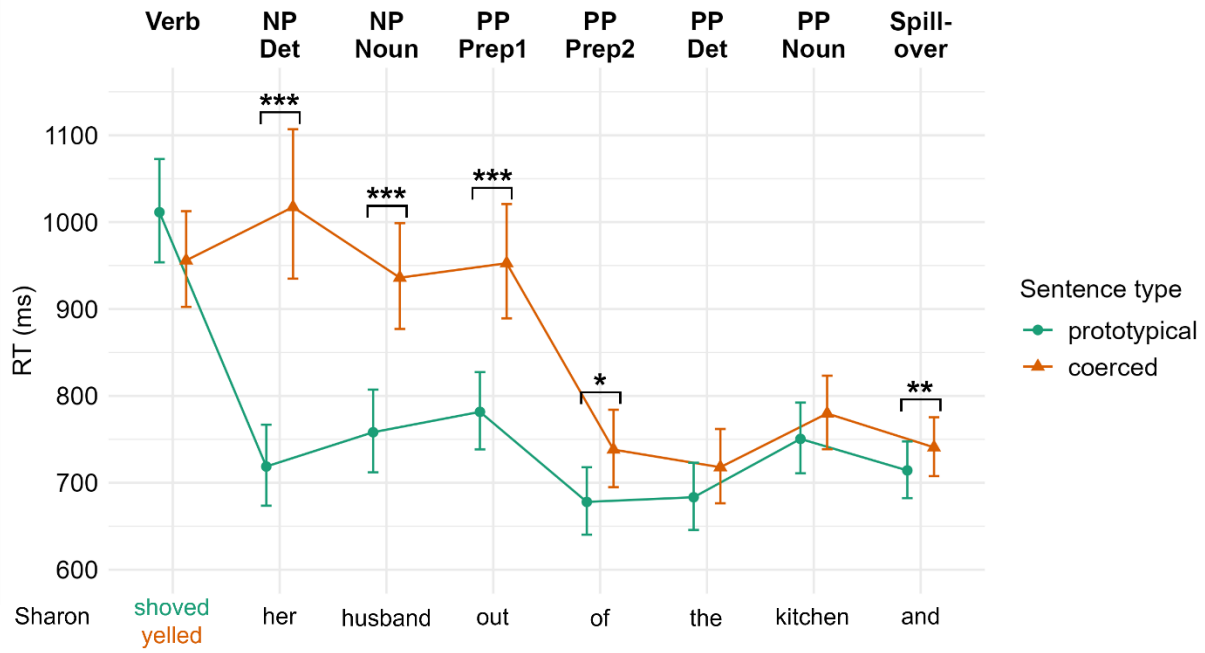
Main Analysis of Response Times

Starting with the main variable of interest, the effect of sentence type on RTs varied depending on the word region. These differences are illustrated in

Figure 2, with the logarithmic values back-transformed into their original RT metric. At the verb, there was no statistically significant main effect of sentence type, suggesting that RTs did not differ between prototypical and coerced sentences (β = -0.053, *SE* = 0.035, *t* = -1.52, *p* = .14). At the following four words, however, RTs differed significantly, with the differences gradually decreasing in magnitude. Relative to prototypical sentences, responses to coerced sentences were estimated to be 299 ms slower at the determiner of the NP (β = 0.347, *SE* = 0.045, *t* = 7.75, *p* = <.001), 178 ms slower at the noun of the NP (β = 0.211, *SE* = 0.039, *t* = 5.38, *p* = <.001), 171 ms slower at the first constituent of the preposition of the PP (e.g., *out* in *out of*; β = 0.198, *SE* = 0.037, *t* = 5.35, *p* = <.001), and 60 ms slower at the second constituent of the preposition (e.g., *of* in *out of*, but note that not all stimuli had two-part prepositions; β = 0.085, *SE* = 0.032, *t* = 2.66, *p* = .01). Subsequently, there were no statistically significant differences between sentence types at the determiner (β = 0.049, *SE* = 0.034, *t* = 1.45, *p* = .15) and the noun of the PP (β = 0.038, *SE* = 0.033, *t* = 1.18, *p* = .25). Finally, at the spillover word, responses to coerced sentences were estimated to be 27 ms slower than those to prototypical sentences (β = 0.036, *SE* = 0.013, *t* = 2.78, *p* = .008).

Figure 2

Estimated RTs by Sentence Type at Each Word in Experiment 1A



Note. Error bars represent 95% confidence intervals. Significance thresholds: *$p < .05$; **$p < .01$; ***$p < .001$. Note that only some stimuli contained a second constituent of the preposition (Prep2).

In terms of covariates, the model at the verb region indicated that RTs decreased as the verb's log frequency increased (β = -0.036, SE = 0.011, t = -3.27, p = .002), while verb length did not have a significant effect (β = 0.004, SE = 0.019, t = 0.23, p = .82). More importantly, we found a significant main effect of trial number at all word regions, suggesting that, averaging across both sentence types, participants' responses became faster as the experiment progressed (all ps < .001). This was further qualified by an interaction between trial number and sentence type at three word regions: the NP noun (β = -0.006, SE = 0.002, t = -3.61, p < .001), the first preposition constituent of the PP (β = -0.004, SE = 0.002, t = -2.15, p = .03), and the determiner

of the PP (β = -0.004, *SE* = 0.001, t = -2.49, p = .01). This indicates that, at these regions, the decrease in RTs over the course of the experiment was stronger for coerced than for prototypical sentences.

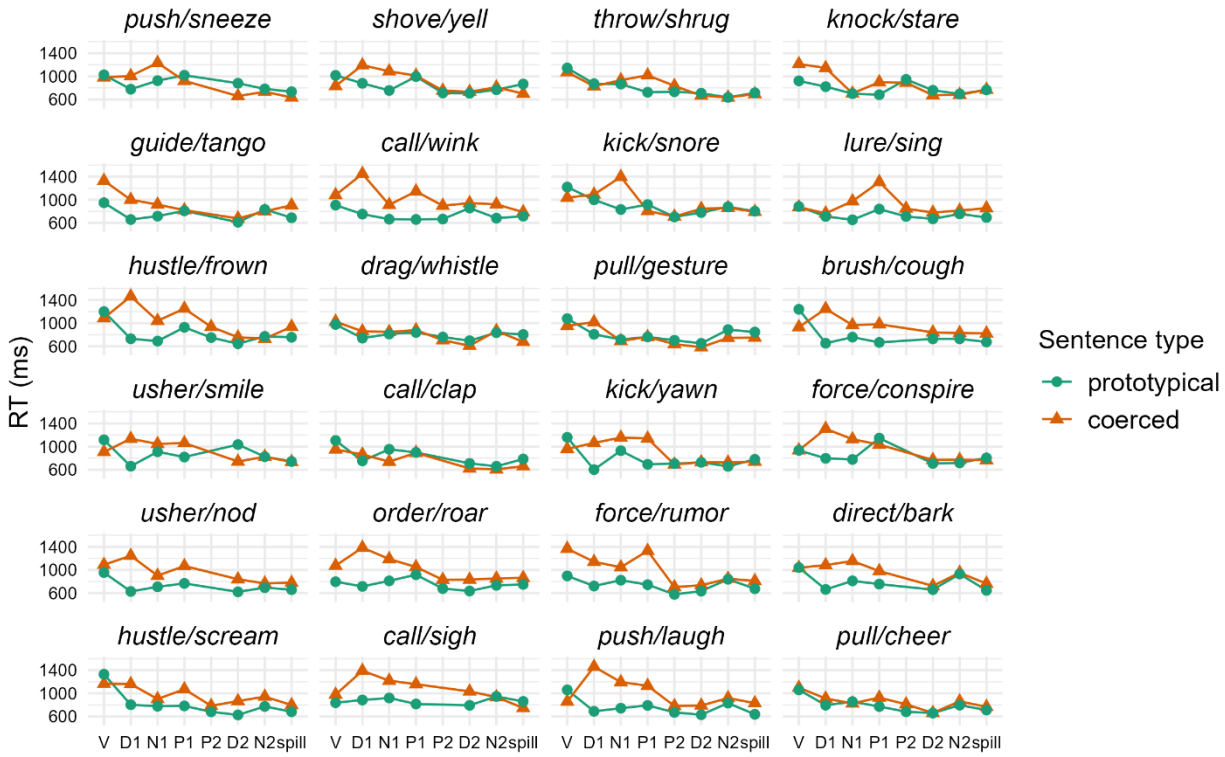
Item-Level Effects on Response Times

We also examined RTs individually by item to determine how consistently the effects emerged across our stimuli.

Figure 3 shows the descriptive results for all 24 items, which are labeled by their prototypical and coerced verb. Visual inspection of the diagram suggests that most items followed the overall pattern, with RTs for coerced sentences being higher than for prototypical sentences at the NP, and these differences gradually disappearing at the PP. Overall, these results suggest that the above-mentioned differences between sentence types generalized to a large proportion of our materials. Nevertheless, the diagram also reveals variability in the response patterns to individual items. Some of this additional variance may be explained by differences in the type of NP violations that occurred in our coerced sentences, as we will explore next.

Figure 3

Observed Item-Level RTs by Sentence Type in Experiment 1A



Note. V = verb; D = determiner; N = noun; P = preposition; spill = spillover word. Note that only some stimuli contained a second constituent of the preposition (P2).

Effects of Violation Type

As discussed earlier (see Processing of Valency Coercion), our coerced sentences displayed two different types of violations at the NP: transitivity and animacy violations. To examine whether this difference affected our results, we compared our models from the main analysis at each word (starting from the NP) with a new set of models in which sentence type was sum-coded as a three-level variable (prototypical vs. coerced with transitivity violation vs. coerced with animacy violation), using likelihood ratio comparisons. Coerced sentences were classified based on the typical argument-linking profiles of their verbs, as determined by the

corpus analyses presented earlier. The only region at which the distinction between these two violation types explained additional variance compared with the main model was the NP determiner ($\chi^2(5) = 54.48$, $p < .001$; all other ps $> .10$). Post-hoc comparisons with the package *emmeans* (Lenth, 2023) showed that responses at this region were slower for coerced sentences with transitivity violations than for coerced sentences with animacy violations ($\beta = -0.184$, $SE = 0.059$, $t = -3.11$, $p = .008$). Compared with prototypical controls, responses were slowed down by both transitivity violations ($\beta = 0.428$, $SE = 0.051$, $t = 8.46$, $p < .001$) and animacy violations ($\beta = 0.243$, $SE = 0.050$, $t = 4.86$, $p < .001$).

Effects of Norming Ratings

We also examined to what extent our results for RTs may have been driven by the sentence acceptability, creativity, and plausibility ratings that we obtained during norming. We did not include these norming variables in our main analysis because they are, by hypothesis, correlated with our critical manipulation of sentence type and would have likely obscured its effects (note that we used the norming ratings to validate our manipulation). Here, we tested their effect by adding them individually to our main model (to reduce collinearity) and examining whether the norming ratings explained any additional variance that was not already accounted for by our main predictor sentence type. Acceptability influenced RTs at only one region, the noun of the NP ($\beta = -0.063$, $SE = 0.028$, $t = -2.27$, $p = .03$), suggesting that less acceptable sentences were read more slowly. When acceptability was included at this region, the effect of sentence type was no longer statistically significant. Creativity marginally influenced RTs at the second preposition constituent of the PP ($\beta = 0.045$, $SE = 0.022$, $t = 2.03$, $p = .06$), indicating that more creative sentences were read more slowly, and also leading to the disappearance of the sentence type effect at this region. Finally, plausibility influenced RTs at the noun of the NP ($\beta =$

-0.077, *SE* = 0.025, *t* = -3.13, *p* = .003) and the first preposition constituent of the PP (β = -0.063, *SE* = 0.024, *t* = -2.66, *p* = .01), suggesting that less plausible sentences were read more slowly. However, sentence type still had a statistically significant effect in the models at both regions.

Accuracy

Finally, we analyzed participants' accuracy in the maze task, but, as noted above, most of the statistical models at the critical words did not converge. This is likely due to the fact that accuracy scores hardly differed between conditions and were almost at ceiling (97% or higher at all sentence regions for both sentence types). The two models that did converge showed no main effect of sentence type (both *p*s > .10). There were, however, a marginally significant interaction between sentence type and trial number at the second preposition constituent of the PP (β = 0.124, *SE* = 0.070, *z* = 1.77, *p* = .08), suggesting that accuracy increased over the course of the experiment for coerced relative to prototypical sentences, and a significant interaction in the opposite direction at the spillover word (β = -0.180, *SE* = 0.078, *z* = -2.31, *p* = .02), suggesting that accuracy decreased for coerced relative to prototypical sentences.

Discussion

Word-by-word RTs in the maze task demonstrated that participants experienced significant processing difficulty while reading coerced sentences, as compared with prototypical controls. This additional processing cost emerged immediately following the verb, at the determiner of the postverbal NP (e.g., *sneezed <u>his</u> napkin*). Differences in RTs persisted but gradually decreased in magnitude until after participants had read the preposition of the locative PP (e.g., *<u>off</u> the table*). Across the remainder of the PP, processing of coerced and prototypical sentences was statistically indistinguishable, although a small difference at the spillover word suggested that coerced sentences caused some persistent, albeit minor, difficulty. Together, these

findings suggest that coerced sentences give rise to a temporary anomaly at the NP, which is, however, rapidly resolved at the start of the PP. There was also evidence that participants' processing of the coerced structures was gradually facilitated as the experiment progressed, as indicated by significant interactions between sentence type and trial number at several word regions.

Further analyses revealed an effect of NP violation type at the NP determiner, where transitivity violations (e.g., *sneezed the napkin …*) induced more difficulty than animacy violations (e.g., *yelled her husband …*). This is in line with our predictions because the determiner directly conflicts with the argument-linking requirements of prototypically intransitive verbs. Interestingly, coerced sentences with animacy violations still gave rise to a processing cost at the NP determiner compared with prototypical sentences, even though their verbs should, in principle, allow inanimate NPs. However, as our corpus norms show (see Appendix B), many of these verbs are more frequently used intransitively, so the presence of the determiner may have conflicted with comprehenders' expectations concerning the verbs' preferred argument structure. Overall, this suggests that semantic properties of the object phrase—in particular, its animacy—had a lesser and more localized effect on responses than the structural properties of the coerced verb-argument composition, which affected processing across several postverbal word regions.

Including our by-item norming ratings in the analyses explained only limited additional variance beyond what was accounted for by the difference between sentence types. Specifically, sentence acceptability and creativity outperformed sentence type as a predictor of RT at only one word region each, with the effects going in the expected direction (slower responses to less acceptable and more creative sentences). Plausibility significantly improved the models at two

regions, but sentence type nevertheless made an independent contribution in each case. This suggests that the difference between coerced and prototypical sentences captures a significant amount of variance in the data, and that it is not primarily driven by the potentially confounding effects of plausibility (see our discussion in Materials and Norming).

Finally, participants' accuracy in the maze task was consistently very high and varied only marginally over the course of the experiment. Coerced sentences, despite being temporarily anomalous, were still clearly preferred to the contextually inappropriate maze distractors. Tentatively, we take this as a sign that comprehenders may be rather flexible in entertaining possible, even non-canonical, relations between grammatical arguments.

While Experiment 1A provided fine-grained evidence about the comprehension of coerced sentences, it is essential to establish the empirical robustness of these patterns. This is especially important given that the linguistic structures examined here have previously not been tested in an on-line comprehension setting, and that our predictions had consequently captured broader processing trends rather than, for instance, specifying at which exact word region the difference between coerced and prototypical sentences would disappear. We therefore conducted an exact replication of the experiment.

Experiment 1B

Experiment 1B was a preregistered exact replication of the maze task Experiment 1A, aimed to assess the robustness of the previously observed findings.

Power Analysis

In order to select an appropriate sample size, we performed a simulated-based power analysis using the results of Experiment 1A. Specifically, we determined how many participants were required to observe, with at least 80% power, the fixed effects of sentence type at the five

sentence regions at which Experiment 1A had yielded statistically significant differences. To do so, we used the powerCurve function from the package *simR* (Green & MacLeod, 2016) to conduct 5,000 simulations of the main statistical models at each region and at varying sample sizes. We did not assess power at the verb region and the later regions of the PP because, by hypothesis, an effect of sentence type is not necessarily expected at these regions. Our power analysis (see the supplemental materials for details) indicated that 120 participants were needed to achieve sufficient power at all relevant regions.

Method

Participant recruitment, materials, procedure, data preprocessing, and analyses, were identical to Experiment 1A and preregistered at https://doi.org/10.17605/OSF.IO/KTG7B. After replacing one participant who did not meet the accuracy thresholds and excluding RTs at words with incorrect maze choices as well as outlier RTs (3.5% of the data), we ended up with 10,599 datapoints (RT) and 10,980 datapoints (maze accuracy) across eight words regions from 120 participants. In our statistical models, maximally converging random effects included random intercepts for participants and items at all regions as well as a random by-participant slope for sentence type at the NP determiner, the NP noun, and the first constituent of the preposition of the PP.

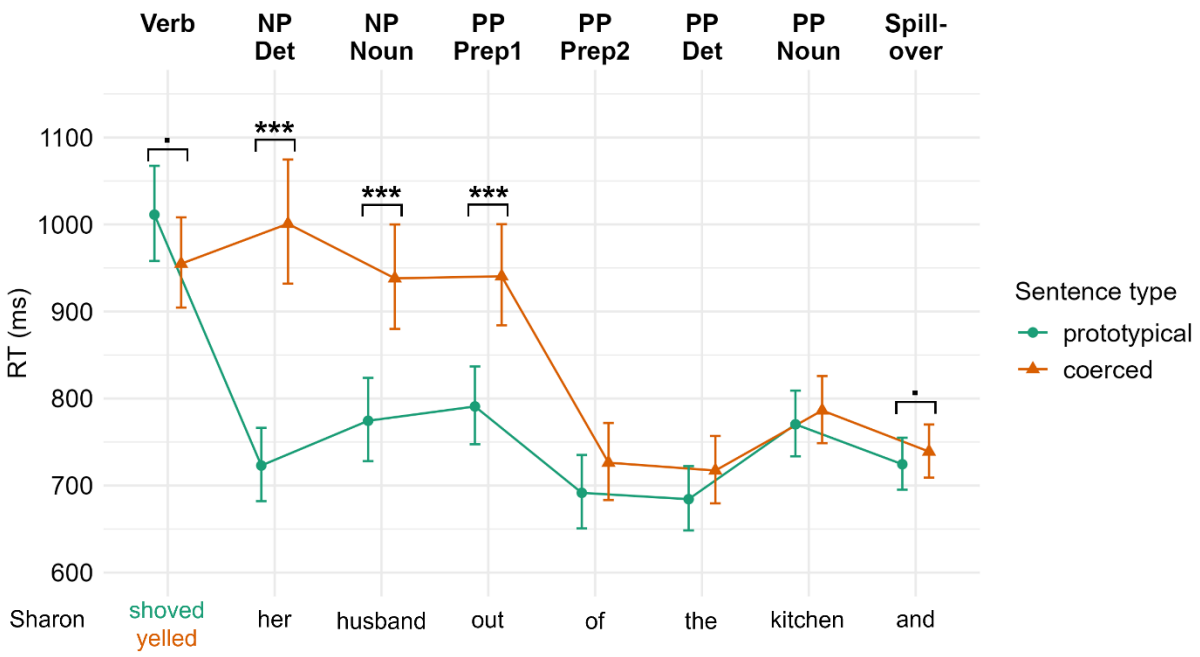
Results

Given that the results were highly similar to Experiment 1A, we briefly summarize them here, including any discrepancies from the original experiment. The key effects of sentence type on (back-transformed) RT are illustrated in **Figure 4**, which largely parallels the earlier

Figure 2. Detailed model outputs and additional (e.g., item-level) plots are available in the supplemental materials. As in Experiment 1A, responses to coerced sentences were slower than for prototypical sentences at the NP determiner (β = 0.325, *SE* = 0.037, *t* = 8.83, $p < .001$), the NP noun (β = 0.192, *SE* = 0.038, *t* = 5.06, $p < .001$), the first preposition constituent of the PP (β = 0.173, *SE* = 0.033, *t* = 5.18, $p < .001$), and (albeit only at marginal significance) the spillover word (β = 0.020, *SE* = 0.011, *t* = 1.8, *p* = .08). The magnitudes of these effects (except at the spillover word) were comparable to Experiment 1A, differing by less than 15% from the original back-transformed estimates. In contrast to Experiment 1A, we found no difference between sentence types at the second preposition constituent of the PP (β = 0.049, *SE* = 0.035, *t* = 1.38, *p* = .18). Instead, a marginally significant trend emerged at the verb, suggesting that participants responded faster to coerced sentences than to prototypical sentences (β = -0.057, *SE* = 0.033, *t* = -1.74, *p* = .09). This effect, which was not observed in Experiment 1A, might reflect item-specific lexical characteristics.

Figure 4

Estimated RTs by Sentence Type at Each Word in Experiment 1B



Note. Error bars represent 95% confidence intervals. Significance thresholds: . $p < 0.1$; *$p < .05$; **$p < .01$; ***$p < .001$. Note that only some stimuli contained a second constituent of the preposition (Prep2).

Similar to Experiment 1A, we also found main effects of trial number at all regions (all *p*s < .001) except for the PP noun, suggesting that RTs decreased over the course of the experiment. However, in contrast to Experiment 1A, no interaction between sentence type and trial number emerged in any of our models (all *p*s > .10). As in Experiment 1A, additional analyses showed that NP violation type affected RTs only at the NP determiner ($\chi^2(5) = 11.16$, $p = .004$), with responses being slower for coerced sentences with transitivity violations than for coerced sentences with animacy violations ($\beta = -0.154$, $SE = 0.045$, $t = -3.38$, $p = .004$), which in turn yielded slower responses than prototypical sentences ($\beta = 0.237$, $SE = 0.041$, $t = 5.82$, $p < .001$). Similar to Experiment 1A, there were only few cases in which our models were improved by adding the norming ratings for sentence acceptability, creativity, and plausibility.

Specifically, lower acceptability (β = -0.075, SE = 0.027, t = -2.76, p = .008) led to slower responses at the NP noun, and lower plausibility led to slower responses at the NP noun (β = -0.079, SE = 0.025, t = -3.19, p = .003) and the PP noun (β = -0.041, SE = 0.034, t = -2.52, p = .02). In these cases, sentence type no longer had a statistically significant effect. Meanwhile, creativity did not improve any of the models.

Finally, the models of maze accuracy converged at all regions except for the verb and the first preposition constituent of the PP. As in Experiment 1A, none of the models indicated a main effect of sentence type (all ps > .10). There were, however, marginally significant interactions between sentence type and trial number at the NP determiner (β = 0.088, SE = 0.051, z = 1.70, p = .09) and the NP noun (β = 0.182, SE = 0.103, z = 1.77, p = .08), suggesting that accuracy increased over the course of the experiment for coerced relative to prototypical sentences.

Discussion

The results of the preregistered replication closely mirror the key effects observed in Experiment 1A, including the substantial slowdown in RT following coerced verbs and the rapid alleviation of this difficulty once comprehenders had read the locative preposition. In the subset of sentences that contained a second preposition constituent (e.g., *out of the kitchen*), the difference between sentence types was no longer statistically significant at this word region, but this effect had already been quite small in Experiment 1A and does not play a major role in the overall interpretation of our results.

Experiment 1B also replicated the effect of NP violation type at the NP determiner, where coerced sentences with transitivity violations induced more processing difficulty than those with animacy violations. Similarly, as in Experiment 1A, including sentence acceptability, creativity, and plausibility only improved a few of our models, suggesting that these norming

variables explained little additional variance beyond the effect of sentence type. Finally, the fact that Experiment 1B only yielded main effects of trial number, but no interaction between trial and sentence type (as in Experiment 1A), suggests that participants' responses to coerced sentences did not selectively speed up over the course of the experiment. The reasons for the absence of an interaction are not fully clear but might partly relate to participant-specific response behaviors. Finally, as in Experiment 1A, maze accuracy was always close to ceiling and showed little effect of the sentence type manipulation.

While Experiments 1A and 1B provided robust evidence of the rapid emergence and subsequent resolution of processing difficulty in coerced sentences, they nevertheless leave several questions unanswered. First, due to the nature of the maze task, we were unable to include an additional control condition of anomalous sentences. It remains to be seen whether and how the processing of coerced sentences differs from these fully ungrammatical controls, which could shed light on the specific nature of grammatical creativity as opposed to uninterpretable grammatical deviation. Second, the stepwise word choices in the maze task differ from participants' natural mode of reading, thus potentially reducing the ecological validity of the method (Forster et al., 2009; Witzel et al., 2012). Third, the maze task does not allow participants to reread previous words in the sentence, even though this may be an important strategy by which readers repair the unusual structure of coerced sentences. To address these limitations, we conducted a second experiment in which participants' eye movements were recorded during normal reading, where they could look back to earlier sentence regions.

Experiment 2

Experiment 2 used eye-tracking to investigate the comprehension of coerced sentences, compared with prototypical and anomalous controls, in a more ecologically valid reading setting.

Eye tracking provides rich evidence both of the temporally "early" stages of processing as participants read a sentence region for the first time, and "later" processing stages during which they may integrate or reanalyze elements in relation to the preceding words (Clifton et al., 2007; Vasishth et al., 2013).

Participants

A total of 55 participants were recruited from the Concordia University community. All of them were self-reported native speakers of English, having acquired English before the age of 5 and using it as (one of) their dominant language(s). Participants all had normal or corrected-to-normal vision. They participated either for course credits, monetary compensation ($10.50), or as unpaid volunteers.

Materials

We used the same materials as in Experiment 1 but with an additional control condition of anomalous sentences, as discussed and illustrated above (see Materials and Norming). The stimuli consisted of 24 text passages, each of which contained three different target versions (prototypical, coerced, anomalous), for a total of 72 sentences.

We distributed the stimuli over three lists so that each participant only saw one version of each target sentence (i.e., 24 items). None of the target verbs occurred more than twice on the same list. We also added 24 filler items that were similar in length to the critical items, although they consisted of four rather than three sentences. These passages belonged to another study that investigated different linguistic structures.

Apparatus and Procedure

The stimuli were presented in white font on a black background, using a ViewSonic 19" CRT monitor (model G90fb, 1,020×768-pixel resolution, 100-Hz refresh rate). Participants were

seated 60 cm in front of the screen so that 1° of visual angle corresponded to approximately three to four characters. We used Experiment builder (Version 2.3) to present the stimuli and record the responses. While participants read the sentences, their eye positions were recorded (based on the right eye) using an EyeLink II head-mounted eye tracker (SR Research, Ottawa, Ontario) with a sampling resolution of 500 Hz (viewing was binocular).

Participants read the text passages sentence-by-sentence, with each sentence presented on a separate screen. They pressed a button to initiate the next sentence. All trials began with drift correction. At the start of every sentence, a gaze-contingent fixation cross appeared on the left of the screen. Once participants had fixated the cross for 120 ms, the sentence appeared. After 25% of trials, participants answered a yes/no comprehension question about information provided in the context sentences (but never the target sentence), to which they responded by a button press.

Data Preprocessing and Analysis

Using Data Viewer (SR Research), we first removed all fixations that occurred before participants started reading the sentences. We then manually applied a vertical drift correction—that is, we corrected eye positions that had been recorded above or below the words due to calibration issues, but which clearly formed part of a linear scan path that followed the sentence across the screen. Meanwhile, we removed fixations that vertically deviated from this scan path, given that these fixations most likely did not occur on the sentence.

All subsequent analyses were conducted in R. We first excluded eight participants who answered less than 70% of comprehension questions correctly, leaving us with data from 47 participants. We then analyzed four eye-tracking measures that have also been used in previous studies of sentences with NP violations (e.g., Staub, 2007; Warren et al., 2015; Warren & McConnell, 2007), defined as follows. (a) *First-pass time* is the sum of all fixations in a region

before readers exit it for the first time, excluding cases in which they have previously fixated a region further to the right. (b) *First-pass regressions out* is the percentage of trials in which readers exit a sentence region to the left after first-pass reading to look back at a previous region. First-pass time and regressions out are regarded as early measures of processing that may reflect lexical access and initial integration with the context (Clifton et al., 2007; Rayner, 2009; Vasishth et al., 2013). (c) *Regression-path time* is the sum of all fixations in a region, including regressive movements, from first entering it until leaving it to the right. This measure is thought to index both early and slightly later processing because it includes regression durations, which may reflect the cost of overcoming integration difficulty (Clifton et al., 2007). Finally, (d) *total time* is the sum of all fixations in a region. This measure reflects the total amount of processing, including temporally late stages in which a region is re-read.

Prior to analysis, we removed all observations where participants had not fixated on a given region during first-pass reading, meaning that any regressions to this region would not be indicative of re-reading (3.9%). As in Experiment 1, we log-transformed the reading time measures (first-pass, regression-path, and total time) to fulfill normality assumptions. We then fitted separate models for each outcome measure at each of the four regions of interest: the verb, the NP, the PP, and the following two words (always *and* plus a verb), which we treated as a potential spillover region. Using the *lmerTest* package (Kuznetsova et al., 2017), we used linear mixed regression for first-pass, regression-path, and total times, and logistic mixed regression for regressions out. The predictor variable of interest in all models was sentence type, sum-coded as prototypical (1,0), coerced (0,1), anomalous (-1,-1). As in Experiment 1, we also included trial number (centered around the mean) as well as the interaction between trial number and sentence type in all models. Verb length and the verbs' log-transformed frequency were further added as

covariates to the models at the verb region. Maximal random effects included random intercepts for subject and items in all models, and a random by-participant slope for sentence type in the models of first-pass times at the verb and spillover regions, and the model of regression-path times at the verb. For all other models, these random slopes either did not converge or produced singular fits.

To assess the overall effect of sentence type, we compared each model with a null model that did not include sentence type, using maximum likelihood tests. If this was significant, we conducted post-hoc comparisons between the three sentence types with the package *emmeans* (Lenth, 2023), using Tukey adjustments for multiple comparisons. For verb length and frequency, *p*-values were computed with the Satterthwaite approximation for degrees of freedom (Kuznetsova et al., 2017). For trial number, statistical significance was established via stepwise maximum-likelihood comparisons with models that included no interaction between trial number and sentence type, or no main effect of trial number. The full outputs of all statistical models are available in the supplemental materials.

Results

Main Analysis

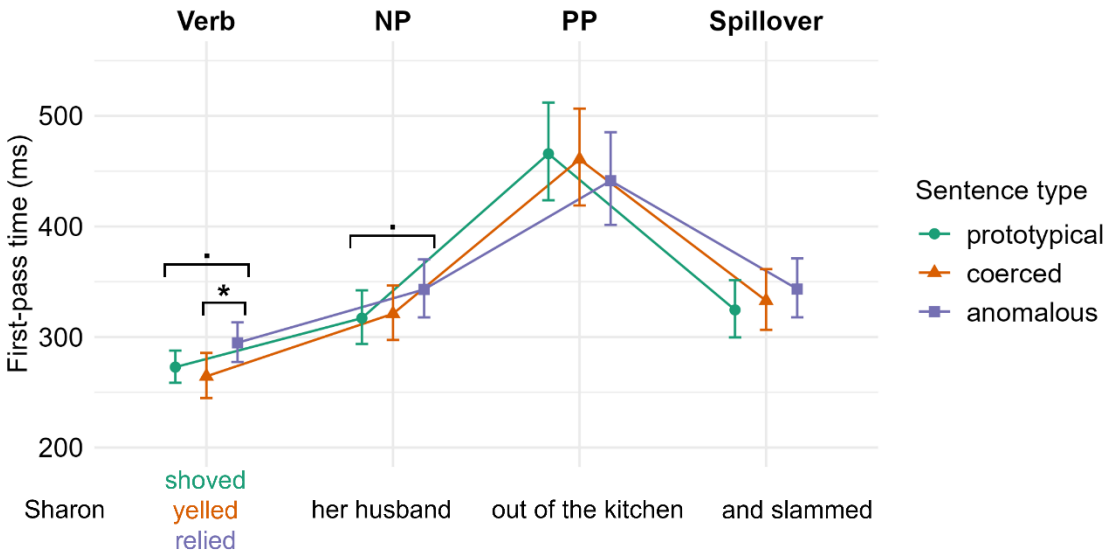
We start by reporting statistically the significant effects of our critical sentence type manipulation on each eye-tracking measure.

First-Pass Time. Model estimates are illustrated in

Figure 5, where the logarithmic values are back-transformed into their original reading time metric. Comparisons with null models indicated that sentence type affected regression-path times at the verb region ($\chi^2(4) = 10.54$, $p = .03$) and at the NP region, though only at marginal significance ($\chi^2(4) = 9.39$, $p = .052$). Post-hoc comparisons showed that, at the verb region, coerced sentences were read 30 ms faster than anomalous sentences ($\beta = -0.109$, $SE = 0.039$, $z = -2.77$, $p = .02$), and prototypical sentences were read 22 ms faster than anomalous sentences, though only marginally so ($\beta = -0.078$, $SE = 0.033$, $z = -2.33$, $p = .07$). At the NP region, there was only one marginally significant contrast, with prototypical sentences being read 26 ms faster than anomalous sentences ($\beta = -0.079$, $SE = 0.037$, $z = -2.14$, $p = .09$).

Figure 5

Model-Estimated First-pass time by Sentence Type at Each Region in Experiment 2

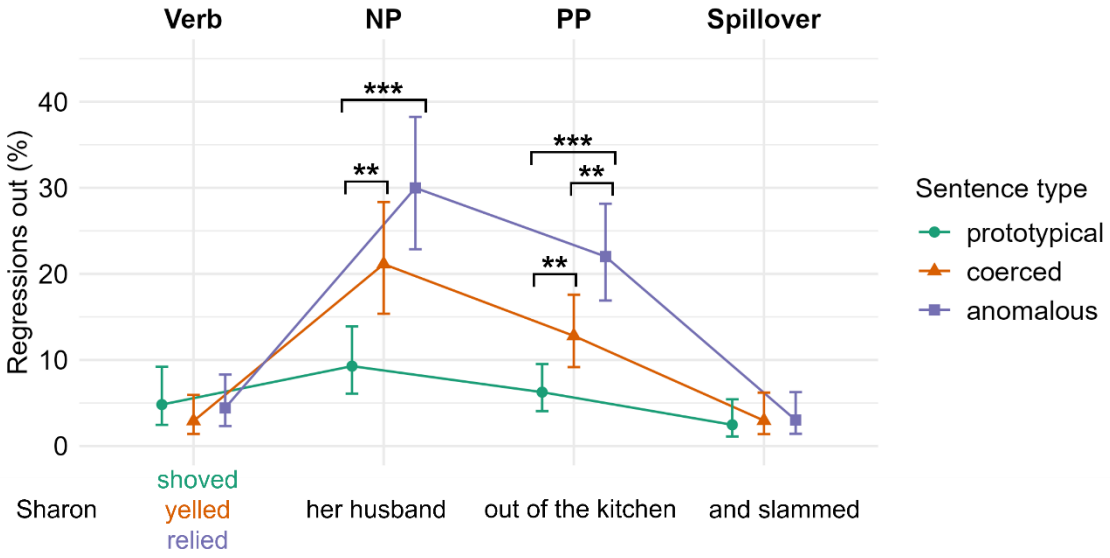


Note. Error bars represent 95% confidence intervals. Significance thresholds: . $p < 0.1$; *$p < .05$.

Regressions out. Model estimates are illustrated in **Figure 6**. Comparisons with null models indicated that sentence type affected regressions out at the NP region ($\chi^2(4) = 26.71$, $p < .001$) and at the PP region ($\chi^2(4) = 31.71$, $p < .001$). Post-hoc comparisons showed that, at the NP region, the percentage of regression was 12% greater in coerced sentences than in prototypical sentences ($\beta = -0.96$, $SE = 0.28$, $z = -3.49$, $p = .001$) and 21% greater in anomalous sentences than in prototypical sentences ($\beta = -1.43$, $SE = 0.27$, $z = -5.25$, $p < .001$). At the PP region, the percentage of regressions was 7% greater in coerced sentences than in prototypical sentences ($\beta = -0.80$, $SE = 0.27$, $z = -3.01$, $p = .007$), 9% greater in anomalous sentences than in coerced sentences ($\beta = -0.66$, $SE = 0.22$, $z = -3.06$, $p = .006$), and 16% greater in anomalous sentences than in prototypical sentences ($\beta = -1.46$, $SE = 0.25$, $z = -5.75$, $p < .001$).

Figure 6

Model-Estimated Regressions Out by Sentence Type at Each Region in Experiment 2



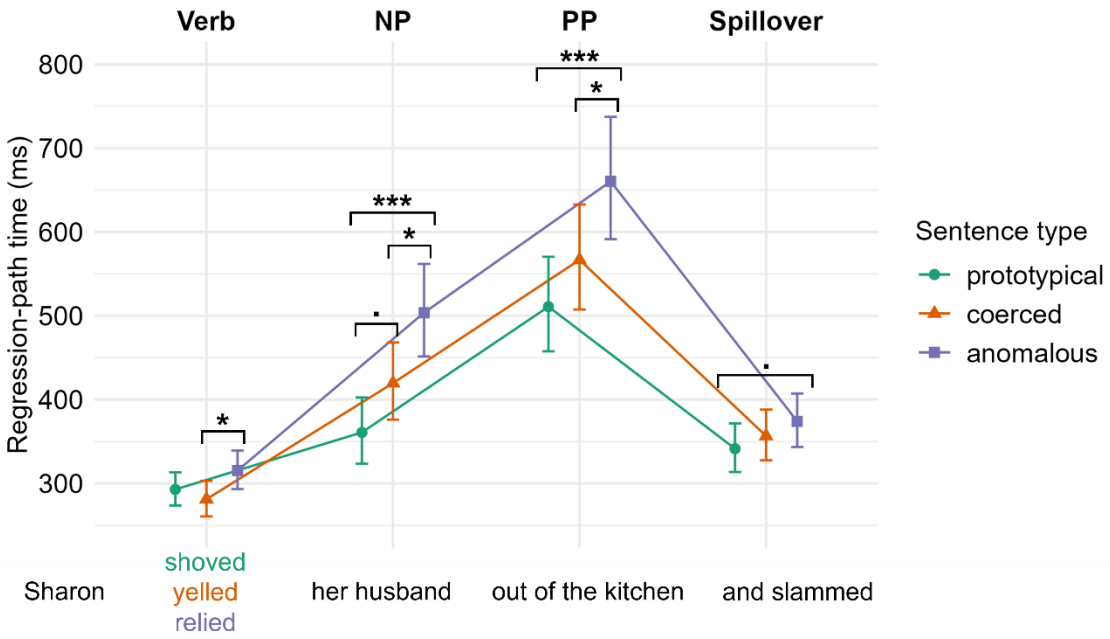
Note. Error bars represent 95% confidence intervals. Significance thresholds: *$p < .05$; **$p < .01$; ***$p < .001$.

Regression-Path Time. Model estimates are illustrated in **Figure 7**. Comparisons with null models indicated that sentence type affected regression-path times at the verb region, though only marginally ($\chi^2(4) = 8.57$, $p = .07$), at the NP region ($\chi^2(4) = 26.50$, $p < .001$), and at the PP region ($\chi^2(4) = 21.44$, $p < .001$). Post-hoc comparisons showed that, at the verb region, coerced sentences were read 34 ms faster than anomalous sentences ($\beta = -0.115$, $SE = 0.043$, $z = -2.68$, $p = .03$). At the NP region, coerced sentences were read 59 ms more slowly than prototypical sentences, though only at marginal significance ($\beta = -0.151$, $SE = 0.063$, $z = -2.38$, $p = .052$) and 84 ms faster than anomalous sentences ($\beta = -0.183$, $SE = 0.063$, $z = -2.88$, $p = .01$), while prototypical sentences were read 143 ms faster than anomalous sentences ($\beta = -0.333$, $SE = 0.063$, $z = -5.27$, $p < .001$). At the PP region, coerced sentences were read 94 ms faster than

anomalous sentences (β = -0.153, SE = 0.054, z = -2.81, p = .02), and prototypical sentences were read 149 ms faster than anomalous sentences (β = -0.257, SE = 0.054, z = -4.73, p < .001). Finally, at the spillover region, there was only one marginally significant contrast, indicating that prototypical sentences were read 33 ms faster than anomalous sentences (β = -0.091, SE = 0.041, z = -2.25, p = .07).

Figure 7

Model-Estimated Regression-Path Time by Sentence Type at Each Region in Experiment 2



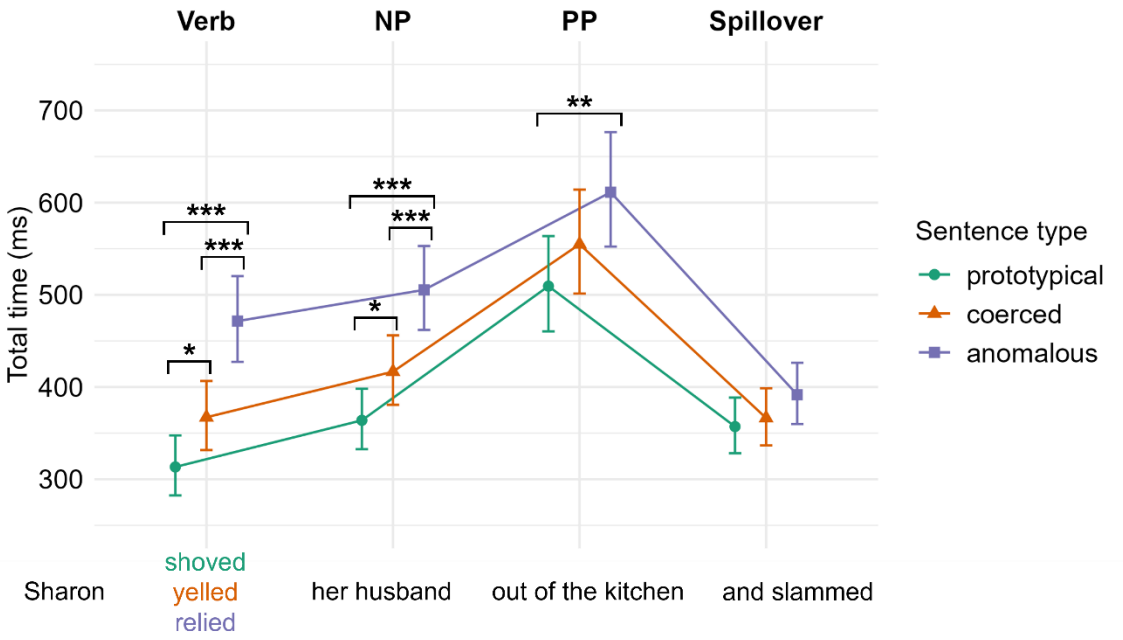
Note. Error bars represent 95% confidence intervals. Significance thresholds: . p < 0.1; *p < .05; **p < .01; ***p < .001.

Total Time. Model estimates for total times are illustrated in **Figure 8**. Comparisons with null models indicated that sentence type had an effect at the verb region ($\chi^2(4) = 40.71$, p < .001), at the NP region ($\chi^2(4) = 49.34$, p < .001), and at the PP region ($\chi^2(4) = 12.98$, p = .01).

Post-hoc comparisons showed that, at the verb region, coerced sentences were read 54 ms more slowly than prototypical sentences (β = -0.159, *SE* = 0.064, z = -2.47, p = .04) and 104 ms faster than anomalous sentences (β = -0.250, *SE* = 0.059, z = -4.26, p < .001), while prototypical sentences were read 158 ms faster than anomalous sentences (β = -0.409, *SE* = 0.060, z = -6.78, p < .001). At the NP region, coerced sentences were read 53 ms more slowly than prototypical sentences (β = -0.135, *SE* = 0.045, z = -3.02, p = .01) and 89 ms faster than anomalous sentences (β = -0.193, *SE* = 0.045, z = -4.30, p < .001), while prototypical sentences were read 142 ms faster than anomalous sentences (β = -0.328, *SE* = 0.045, z = -7.33, p < .001). At the PP region, only one contrast was significant, suggesting that prototypical sentences were read 102 ms faster than anomalous sentences (β = -0.182, *SE* = 0.049, z = -3.70, p = .001).

Figure 8

Model-Estimated Total Time by Sentence Type at Each Region in Experiment 2



Note. Error bars represent 95% confidence intervals. Significance thresholds: *$p < .05$; **$p < .01$; ***$p < .001$.

Covariates in the main models. Across eye-tracking measures, we found no effect of verb length at the verb region. We did, however, observe effects of (log-transformed) verb frequency on three of the four measures, indicating that verbs with higher frequency gave rise to shorter first-pass times, fewer regressions out, and shorter regression-path times (all *p*s < .05). There were also main effects of trial number at all sentence regions, indicating that reading times decreased over the course of the experiment, even though the effects did not emerge consistently across all measures (see the supplemental materials for detailed test results). Interactions between trial number and sentence type were only statistically significant in two cases: for total times at the NP ($\chi^2(2) = 7.84$, $p = .02$) and first-pass times at the PP ($\chi^2(2) = 6.30$, $p = .04$). Post-hoc comparisons revealed that, as the experiment progressed, total times at the NP decreased more for anomalous sentences ($p = .03$) and, marginally so, for coerced sentences ($p = .06$) than for prototypical sentences; and that first-pass times at the PP decreased more for prototypical than anomalous sentences, though also only marginally ($p = 0.055$).

Item-Level Effects

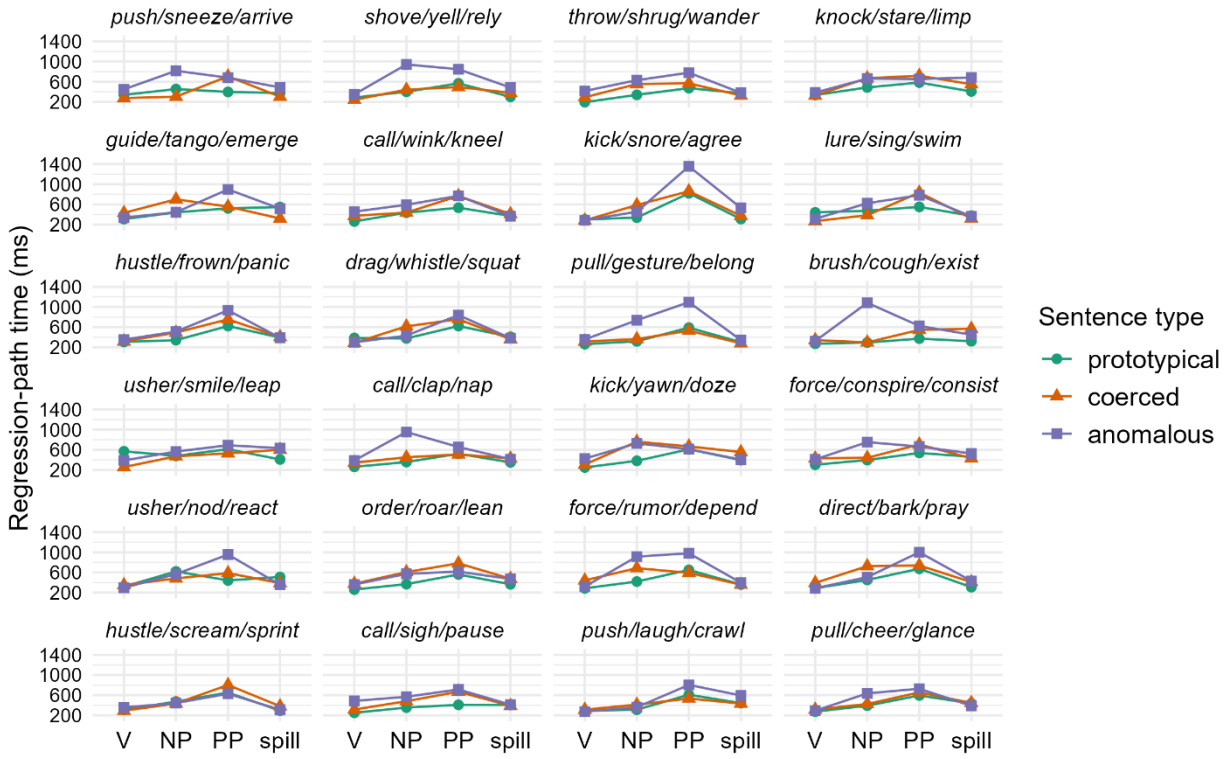
As in Experiment 1, we also examined our results at the level of the individual items. We focus on regression-path times here as they showed a clear pattern in the overall results; item-level results for the other measures are available in the supplemental materials.

Figure 9 depicts the average regression-path times for all 24 items, labeled by their prototypical, coerced, and anomalous verbs. The plot reveals a considerable amount of item-level variation. Nevertheless, many items follow the expected trend in that they show longer regression-path times for coerced sentences than prototypical sentences at the NP (and

sometimes PP) as well as even longer reading times in the anomalous condition. This suggests that the above-mentioned results generalize to a substantial subset of our materials.

Figure 9

Observed Item-Level Regression-Path Times by Sentence Type in Experiment 2



Effects of Violation Type

As in Experiment 1, we investigated whether the type of NP violation in coerced sentences (transitivity vs. animacy violation) affected our results. For this purpose, we fitted new models with a four-level variable sentence type that distinguished between the two types of coerced sentences and compared them with the models from our main analysis. However, we did not find improvements in any of our models at the NP or the subsequent regions (all *p*s > .10).

Effects of Norming Ratings

We also examined to what extent our norming ratings for sentence acceptability, creativity, and plausibility ratings may have driven our results. As in Experiment 1, we added each norming variable separately to our main models and compared them with the original models. We summarize the results here, but detailed test results are available in the supplemental materials. Acceptability ratings improved our models of regressions out at the verb region (p = .04), the NP region (p = .01), and the PP region (p = .04), our model of regression-path time at the NP region (p = .01), and our model of total time at the verb region (p = .02) and, marginally, the NP region (p = .09). In most cases, the effects of sentence type were no longer statistically significant when acceptability was included in the models. Creativity ratings did not improve any of our models, except for a marginal effect on regressions out at the PP (p = .054). Plausibility ratings affected the same regions and eye-tracking measures as acceptability (all ps < .05). In some of these models, sentence type no longer had a statistically significant effect when plausibility was included; but in the models of total times, sentence type still had an effect that was not explained by plausibility.

Discussion

Experiment 2 demonstrated clear differences in eye-movement patterns during the comprehension of coerced, prototypical, and anomalous sentences. Effects in first-pass time were of little interest: While anomalous sentences were read more slowly at the verb and potentially the NP, this is most likely due to the low contextual fit of their verbs. In contrast, regressions out yielded processing differences between all three sentence types, supporting prior evidence that effects in first-pass time and first-pass regressions can trade off with each other (Rayner & Sereno, 1994). At the NP, both coerced and anomalous sentences yielded a greater proportion of

regressions than prototypical sentences, suggesting that the violation of their verbs' prototypical argument-linking patterns induced processing difficulty at a temporally early stage. At the PP, these effects persisted, but crucially, coerced sentences gave rise to significantly fewer regressions than anomalous sentences, a contrast that had not been statistically significant at the NP. This suggests that the PP enables participants to at least partly overcome the processing difficulty in coerced sentences, thus making them less likely to revisit previous sentence regions.

The results for regression-path time, which include the duration of regressive eye movements, suggest that processing of coerced sentences was additionally alleviated at a somewhat later temporal stage. At the PP, coerced sentences no longer differed statistically from prototypical sentences, indicating that even if participants regressed at that point, their regressions were shorter. Finally, total time, when contrasted with regression-path time, sheds light on temporally late re-reading efforts. These results indicate that, especially at the verb region, coerced sentences caused longer re-reading than prototypical sentences but shorter re-reading than anomalous sentences.

Main effects of trial number indicated that reading times and regression probabilities decreased over the course of the experiment, which could be due to participants' growing familiarity with the structures or task, or due to a decrease in attention. In contrast to Experiment 1A, but in line with Experiment 1B, there was limited evidence for an interaction with sentence type, suggesting that these trial effects did not specifically affect coerced sentences. Another difference from Experiments 1A and 1B was that the type of NP violation in coerced sentences (transitivity vs. animacy) did not affect our results. This is unsurprising, given that Experiment 1 had suggested that violation types only had a differential effect at the NP determiner, whereas Experiment 2 measured across the entire NP region.

Including our by-item norming ratings in the analyses had a more substantial impact than in Experiments 1A and 1B. Sentence acceptability was a better predictor of eye-movement patterns at several regions than our sentence type variable. This is expected because sentence types were designed to differ in acceptability, but it additionally suggests that item-level differences in acceptability explained a substantial amount of variance that was not accounted for by the coarse classification into sentence types. Creativity, on the other hand, explained little variance beyond the effect of sentence type. Finally, plausibility was also a good predictor of eye movements. As discussed earlier (see Materials and Norming), this raises the question of whether the differences between prototypical and coerced sentences were driven by structural factors or by their (unintended) difference in plausibility. However, for total time at least, sentence type and plausibility explained partly independent amounts of variance, suggesting that plausibility cannot account for our full set of results. In addition, given that Experiment 1A demonstrated differences between prototypical and coerced sentences that exceeded the effects of plausibility, it seems unlikely that the same contrasts in Experiment 2 were exclusively driven by plausibility.

General Discussion

In two experiments, we investigated how grammatically creative sentences are processed during incremental on-line (i.e., real-time) comprehension. Our results provide, to our knowledge, the first evidence of how readers comprehend creative combinations of verbs and their grammatical arguments in cases of so-called valency coercion. We will first discuss our key findings regarding the time-course of valency coercion processing, before addressing their implications for theories of argument structure and some potential areas for follow-up research.

The Time-Course of Valency Coercion Comprehension

Our experiments demonstrate that the comprehension of coerced sentences in which verbs are combined with non-canonical object and locative goal arguments differs from that of prototypical, structurally unmarked sentences as well as anomalous, fully ungrammatical sentences. Examples of the three sentence types are shown again in (11).

(11) a. *Frank (sneezed / pushed / arrived) his napkin off the table.*

b. *Sharon (yelled / shoved / relied) her husband out of the kitchen.*

Experiment 1A and its preregistered replication in Experiment 1B used the maze task to identify the *locus* of processing effects in prototypical and coerced sentences on a word-by-word basis. Coerced sentences gave rise to immediate difficulty at the determiner and the noun of the postverbal NP. These effects emerged both when the NP followed a prototypically intransitive verb, such as *sneezed* in (11a), and when the NP contained an animate noun but followed a verb that can only select inanimate nouns, such as *yelled* in (11b). This is consistent with prior evidence from reading comprehension showing that transitivity violations (Mitchell, 1987; Staub, 2007; van Gompel & Pickering, 2001) and violations of verbs' selectional restrictions (Paczynski & Kuperberg, 2012; Warren et al., 2015; Warren & McConnell, 2007) produce immediate processing difficulty. Our results also illustrated that transitivity violations led to a higher processing cost at the NP determiner, reflecting the fact that animacy violations were only detected at the noun.

Crucially, Experiments 1A and 1B shed light on how comprehenders subsequently resolved the processing difficulty in coerced sentences, thus illustrating aspects of creative

grammar processing that could not be predicted based on previous studies (e.g., Busso et al., 2021). Specifically, once participants had read the preposition of the locative goal (e.g., *off the table*, *out of the kitchen*), processing in coerced sentences was suddenly alleviated and subsequently matched that of prototypical sentences (except for some minor remaining difficulty at the spillover word). This suggests that comprehenders rapidly resolve the initial combinatorial conflict in coerced sentences. Whereas recent experimental work (Busso et al., 2018, 2020, 2021; Perek & Hilpert, 2014; Yoon, 2016) has shown that instances of valency coercion can be successfully interpreted, the responses in those studies (e.g., acceptability judgments) were only elicited after the end of the sentence and could therefore be driven by some delayed metalinguistic reassessment. In contrast, our results indicate that comprehenders overcome the difficulty of coerced structures through rapid incremental processing before they reach the end of the clause. This aligns with evidence from other types of creative language, such as novel metaphors (Blasko & Connine, 1993; Mashal & Faust, 2009) and novel compounds (Coolen et al., 1993; Gagné, 2002; Libben et al., 1999), suggesting that these expressions are also rapidly comprehended. Blasko and Connine (1993), for instance, found that hearing novel and highly apt metaphors (e.g., *his anger was a blizzard*) facilitates recognition of words related to the metaphorical meaning (e.g., *blinding*) within less than 1 second after the offset of the metaphorical noun. Our results extend these findings from lexical-semantic creativity to the incremental processing of grammatically creative expressions.

Experiment 2, which used eye-tracking to compare all three sentence types (coerced, prototypical, anomalous), provided additional evidence about the *temporal stages* of comprehension at each sentence region. Relative to prototypical controls, coerced and anomalous sentences gave rise to processing difficulty both at temporally early stages, where they led to

increased regressions from the NP and PP regions, and at later stages, where they resulted in longer re-reading, especially of the verb region. This is in line with previous evidence from eye-tracking that transitivity violations (Staub, 2007; van Gompel & Pickering, 2001) and animacy violations (Warren et al., 2015; Warren & McConnell, 2007) disrupt temporally early as well as later stages of processing.

In addition, Experiment 2 provided novel evidence on how eye-movement responses to coerced sentences differ from those to fully anomalous sentences, which have been the focus of work on syntactic and semantic violations. Compared with anomalous controls, coerced sentences gave rise to fewer regressions from the PP region, shorter regression-path times at the NP and PP regions, and less re-reading of the verb (as indicated by total time relative to the other measures). Paralleling the results of Experiment 1, this suggests that language users incrementally re-integrated the unusual verb-argument combinations in coerced sentences, in contrast to the persistent processing cost induced by anomalous sentences. The fact that regression-path times in coerced sentences were shorter at the NP than in anomalous sentences suggests that comprehenders already anticipated a potential coerced interpretation at this early point, perhaps due to prior contextual support, which has been shown to rapidly alleviate effects of implausibility (Filik, 2008; Warren et al., 2008). At the same time, such early resolution efforts can have been only partially successful, given that regression-path times for coerced sentences exceeded those for prototypical controls at the NP, whereas this difference was no longer statistically significant at the PP.

Links with Theories of Argument Structure

The present study investigates a core issue bearing on the interplay between grammatical representation (viz., argument structure) and linguistic creativity—the use of language as “an

instrument of free thought and self-expression" (Chomsky, 1966, p. 29). While we have demonstrated that valency coercion can be a mechanism for the "creative aspect of language use," it remains to be seen how the effects we obtained can be accounted for by extant theories of argument structure. We limit our discussion to a few prominent frameworks; for a broad overview of different approaches, see, for instance, Levin (2018). The theories discussed here include both constructional (also called constructionist or phrasal) approaches, which posit that argument structure is encoded by the clausal structures in which verbs are embedded, and lexical(ist) approaches, which assume that argument structure is encoded by the lexical representations of verbs (for discussion, see Müller & Wechsler, 2014; Williams, 2015).

In the introduction, we outlined a constructional account of valency coercion, based on the framework of Construction Grammar (Boas, 2011; Goldberg, 1995; van Trijp, 2015). In Construction Grammar, grammatical information is assumed to encompass constructional templates that link a specific configuration of syntactic roles to an abstract event meaning. In valency coercion, the construction contributes additional argument roles that are not lexically encoded by the verb. Our experimental results are compatible with this account. Specifically, the structure of sentences such as (11a) may activate comprehenders' knowledge of a "caused-motion construction," which facilitates the construal of the verb *sneeze* as part of a caused motion event. The fact that, in Experiment 1, processing difficulty was alleviated after the locative preposition suggests that the preposition may be a critical syntactic and/or semantic marker that triggers the activation of the caused-motion construction. However, this alleviation only occurred in coerced but not anomalous sentences, thus demonstrating that comprehenders rapidly evaluated the compatibility between the constructional meaning and the lexical semantics

of the verb (e.g., by construing *sneeze* as expressing a manner of motion), and that successful comprehension depends on the interplay of information from both sources.

An alternative phrasal explanation is offered by neo-constructionist approaches, as illustrated, for instance, by Borer's (2003; 2005) exo-skeletal account (see also Cuervo & Roberge, 2012). In this theory, argument structure is exclusively encoded by the structural-syntactic "skeleton" of the sentence. Specifically, the grammar licenses well-formed argument structures on functional-syntactic grounds, whereas verbs (and other substantive, i.e., open-class, lexemes) encode conceptual content but bear no grammatical properties. The plausibility of specific verb-argument combinations, including coerced instances, is determined in an extra-grammatical component – what Borer (2005, p. 9) calls the "making sense" component – that draws on world knowledge to evaluate the fit between the functional interpretation of syntactic structures and the conceptual information provided by the verbs. This theory, too, can account for the processing patterns observed in our experiments: Specifically, the locative preposition, as a functional (i.e., closed-class) item (Borer, 2005, p. 29), may allow comprehenders to parse the syntactic skeleton of the sentence, which is then rapidly evaluated against conceptual and world knowledge, giving rise to a felicitous interpretation.

The constructional approaches discussed so far differ from lexical accounts, which posit that argument-linking patterns are licensed by the lexical representation of verbs (e.g., Grimshaw, 1990; Müller & Wechsler, 2014; Pinker, 1989). In one possible version of this framework (discussed by Levin & Rappaport Hovav, 2005), coerced sentences like (11a) may be accounted for by assuming that language users encode multiple lexical entries for *sneeze*, one that captures its prototypical one-argument sense, 'sneeze(x)', and another one that licenses its use in a three-argument structure, 'cause-to-move-by-sneezing(x,y,z)'. The latter representation

may be infrequently activated and thus less salient, which could explain the temporary processing difficulty in coerced sentences (along the lines of MacDonald et al., 1994). A drawback of this account is that it requires a proliferation of verb entries in the lexicon, which seems "counterintuitive" (Levin & Rappaport Hovav, 2005, p. 190) compared with an approach that generalizes over coerced uses of different verbs. In addition, this lexical account cannot explain how language users comprehend fully novel coerced instances that they have previously not encountered, which may have well been the case for some of the verbs in our experiments that are very rarely attested transitively in corpora (see Materials and Norming).

Other lexical accounts, however, overcome these limitations by assuming that language users do not store additional verb entries that license coerced uses in memory, but that they derive them "on the fly" via lexical rules (Bresnan, 1982; Briscoe & Copestake, 1999; Müller & Wechsler, 2014). In (11a), for instance, the prototypical intransitive representation of *sneeze* serves as input to a lexical rule that returns a three-argument structure, 'cause(sneeze(*x*),move(*x*,*y*,*z*))' (Müller & Wechsler, 2014, pp. 25–26). In contrast to the lexical approach outlined above, however, only the lexical rule, but not its individual outputs, need to be stored in memory, thus avoiding a proliferation of stored verb senses. Since the lexical rule is only triggered when the verb occurs in a specific clausal context, this approach also predicts that coerced instances give rise to processing difficulty until grammatical markers, such as the locative preposition, allow comprehenders to parse the sentence structure.

In sum, our experimental results can be accounted for by several prevalent theories of argument structure. These approaches make use of varying theoretical tools, including constructional templates, lexical rules, and extra-grammatical operations, to explain the occurrence of non-canonical, creative verb-argument combinations. As it stands, processing

evidence of the type presented here cannot adjudicate between the theoretical accounts. Reading times do not shed light on whether the merger of grammatical and conceptual information occurs within a grammatical or an extra-grammatical component, as suggested by Construction Grammar and neo-constructionist approaches, respectively. Nor can they distinguish between the contributions of constructions and lexical rules, which, as Müller and Wechsler (2014, p. 26) note, produce identical "composite" semantic structures. It remains a task for future work to develop experimental manipulations and linking hypotheses that may differentiate between these theories based on specific aspects of processing. Crucially, the present work shows that the creative composition of verbs and grammatically *unlicensed* arguments is obtained in real time, suggesting that, if there are canonical, lexically specified structures, they can be felicitously violated or adjusted during incremental parsing and interpretation.

Limitations and Future Work

Finally, we would like to address some limitations regarding the scope of the present study together with venues for future work on the nature of grammatically creative expressions. First, for the purposes of our case study, we have focused on valency coercion as a subtype of grammatical creativity; and within this phenomenon, we restricted our experiments to coerced instances of a single construction, the English caused-motion construction. Clearly, it is necessary to extend the scope of this investigation, both to other coerced constructions in more diverse languages and also to grammatically creative structures that do not solely rely on the verb's valency.

Second, by embedding the coerced stimuli in naturalistic context passages, our designs lend themselves to further investigation of the role of context in the processing of creative grammar. One relevant question is to what extent participants' comprehension was facilitated by

the contextual support we provided, and how it would be affected by the absence of such cues. In addition, while our coerced stimuli were supported by context, they were nevertheless judged as less plausible than prototypical controls, thus creating a potential confound (see the individual discussions of Experiments 1 and 2). Future work could use targeted manipulations to distinguish more clearly between the contributions of grammatical creativity and plausibility. Further linguistic or situational characteristics of the context could be manipulated to test how malleable the processing effects are in light of participants' prior expectations.

Finally, while our study investigated the time-course of grammatical creativity relying on different measures of reading time, it would be important to extend the present investigation to the neuronal correlates of the "creative aspect of language use," as illustrated by valency coercion. While behavioral data and functional explanations stand on their own, phenomena such as valency coercion constitute a key test case for understanding how different neuroanatomic resources—and, by hypothesis, different sources of information—contribute to the comprehension of grammatical creativity. Verb-argument structures lie at the intersection between lexical and syntactic knowledge, and understanding how they contribute to our productive—and creative—linguistic capacities requires a concerted effort tapping into different methods and levels of analysis.

Conclusion

The present study provides evidence of how non-canonical or "grammatically creative" sentences, which are structurally novel but interpretable, are processed in real time. In particular, our experiments are the first to test the on-line comprehension of valency coercion, where verbs are combined with grammatical arguments with which they usually do not occur—such as when a typically intransitive verb such as *sneeze* takes on two additional grammatical arguments (e.g.,

sneeze the napkin off the table). Using two complementary methods, the "maze" variant of self-paced reading and eye-movement recordings, we obtained detailed evidence about the time-course of parsing at the postverbal arguments. The pattern of response times and eye movements suggests that coerced sentences give rise to immediate processing difficulty after the verb, which is, however, rapidly alleviated and largely disappears before comprehenders reach the end of the clause. We have discussed several possible ways in which our results can be accommodated by prevalent theories of argument structure.

Our main goal has been to investigate the nature of linguistic creativity—how finite linguistic means can yield infinite possibilities—by focusing on valency coercion, a phenomenon that has gained little attention in psycholinguistics. Our results illustrate that argument structure composition is flexible (see also Di Sciullo, 2005), and that deviations from canonical verb-argument-linking are swiftly computed during on-line sentence comprehension. Specifically, comprehenders resolve the temporary anomaly in coerced sentences by incrementally and rapidly integrating information from the verb and its clausal context. We have shown that creative language use goes beyond well-known figurative tropes such as metaphors, extending to grammatically creative verb-argument combinations that, despite their challenges, are naturally accommodated by the comprehension system.

References

Antal, C., & de Almeida, R. G. (2021). Indeterminate and enriched propositions in context linger: Evidence from an eye-tracking false memory paradigm. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.741685

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA

Publications and Communications Board task force report. *The American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arzouan, Y., Goldstein, A., & Faust, M. (2007). Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain Research*, *1160*, 69–81. https://doi.org/10.1016/j.brainres.2007.05.034

Ashby, J., Roncero, C., de Almeida, R. G., & Agauas, S. J. (2018). The early processing of metaphors and similes: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*, *71*(1), 161–168. https://doi.org/10.1080/17470218.2016.1278456

Audring, J., & Booij, G. (2016). Cooperation and coercion. *Linguistics*, *54*(4), 617–637. https://doi.org/10.1515/ling-2016-0012

Baayen, H. R., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.21500/20112084.807

Bader, R., Mecklinger, A., Hoppstädter, M., & Meyer, P. (2010). Recognition memory for one-trial-unitized word pairs: Evidence from event-related potentials. *NeuroImage*, *50*(2), 772–781. https://doi.org/10.1016/j.neuroimage.2009.12.100

Bambini, V., Bertini, C., Schaeken, W., Stella, A., & Di Russo, F. (2016). Disentangling Metaphor from Context: An ERP Study. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00559

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Bergs, A. (2019). What, if anything, is linguistic creativity? *Gestalt Theory*, *41*(2), 173–183. https://doi.org/10.2478/gth-2019-0017

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(2), 295–308. https://doi.org/10.1037//0278-7393.19.2.295

Boas, H. C. (2003). *A constructional approach to resultatives*. CSLI Publications.

Boas, H. C. (2011). Coercion and leaking argument structures in Construction Grammar. *Linguistics*, *49*(6), 1271–1303. https://doi.org/10.1515/ling.2011.036

Borer, H. (2005). *Structuring sense. Vol. 2: The normal course of events*. Oxford University Press.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082. https://doi.org/10.1016/j.jml.2019.104082

Boyce, V., & Levy, R. (2023). A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, *2*(1), 1–34. https://doi.org/10.5070/G6011190

Brennan, J., & Pylkkänen, L. (2008). Processing events: Behavioral and neuromagnetic correlates of aspectual coercion. *Brain and Language*, *106*(2), 132–143. https://doi.org/10.1016/j.bandl.2008.04.003

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. https://doi.org/10.1080/01690961003616840

Bresnan, J. (Ed.). (1982). *The mental representation of grammatical relations*. MIT Press.

Briscoe, T., & Copestake, A. (1999). Lexical rules in constraint based grammars. *Computational Linguistics*, *25*(4), 487–526.

Busso, L., Lenci, A., & Perek, F. (2020). Valency coercion in Italian: An exploratory study. *Constructions and Frames*, *12*(2), 171–205. https://doi.org/10.1075/cf.00039.bus

Busso, L., Pannitto, L., & Lenci, A. (2018). Modelling Italian construction flexibility with distributional semantics: Are constructions enough? In E. Cabrio, A. Mazzei, & F. Tamburini (Eds.), *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)* (pp. 68–72). Accademia University Press.

Busso, L., Perek, F., & Lenci, A. (2021). Constructional associations trump lexical associations in processing valency coercion. *Cognitive Linguistics*, *32*(2), 287–318. https://doi.org/10.1515/cog-2020-0050

Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. *NeuroImage*, *59*(4), 3212–3221. https://doi.org/10.1016/j.neuroimage.2011.11.079

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. Harper & Row.

Christensen, R. H. B. (2022). *Ordinal—Regression models for ordinal data. R package version 2022.11-16*. https://CRAN.R-project.org/package=ordinal

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology*, *70*(7), 1380–1405. https://doi.org/10.1080/17470218.2016.1186200

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–374). Elsevier.

Coolen, R., Van Jaarsveld, H. J., & Schreuder, R. (1993). Processing novel compounds: Evidence for interactive meaning activation of ambiguous nouns. *Memory & Cognition*, *21*(2), 235–246. https://doi.org/10.3758/BF03202736

Cuervo, M. C., & Roberge, Y. (Eds.). (2012). *The end of argument structure*. Emerald.

Davies, M. A. (2008). *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

de Almeida, R. G., & Dwivedi, V. D. (2008). Coercion without lexical decomposition: Type-shifting effects revisited. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, *53*(2–3), 301–326. https://doi.org/10.1353/cjl.0.0026

de Almeida, R. G., Riven, L., Manouilidou, C., Lungu, O., Dwivedi, V. D., Jarema, G., & Gillon, B. (2016). The neuronal correlates of indeterminate sentence comprehension: An fMRI study. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00614

De Swart, H. (1998). Aspect shift and coercion. *Natural Language & Linguistic Theory*, *16*(2), 347–385. https://doi.org/10.1023/A:1005916004600

Delogu, F., Vespignani, F., & Sanford, A. J. (2010). Effects of intensionality on sentence and discourse processing: Evidence from eye-movements. *Journal of Memory and Language*, *62*(4), 352–379. https://doi.org/10.1016/j.jml.2010.02.002

Di Sciullo, A. M. (2005). *Asymmetry in morphology*. MIT Press.

Filik, R. (2008). Contextual override of pragmatic anomalies: Evidence from eye movements. *Cognition*, *106*(2), 1038–1046. https://doi.org/10.1016/j.cognition.2007.04.006

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, *41*(1), 163–171. https://doi.org/10.3758/BRM.41.1.163

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Gagné, C. L. (2002). Lexical and relational influences on the processing of novel compounds. *Brain and Language*, *81*(1), 723–735. https://doi.org/10.1006/brln.2001.2559

Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument structure*. University of Chicago Press.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greenfield, S. B. (1967). Grammar and meaning in poetry. *PMLA*, *82*(5), 377–387. https://doi.org/10.2307/460767

Grimshaw, J. (1990). *Argument structure*. MIT Press.

Harris, J., Pylkkänen, L., McElree, B., & Frisson, S. (2008). The cost of question concealment: Eye-tracking and MEG evidence. *Brain and Language*, *107*(1), 44–61. https://doi.org/10.1016/j.bandl.2007.09.001

Hidalgo-Downing, L. (2015). Metaphor and metonymy. In R. H. Jones (Ed.), *The Routledge handbook of language and creativity* (pp. 107–128). Routledge.

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641–671. https://doi.org/10.1037/bul0000145

Horvat, A. W., Bolognesi, M., Littlemore, J., & Barnden, J. (2022). Comprehension of different types of novel metaphors in monolinguals and multilinguals. *Language and Cognition*, *14*(3), 401–436. https://doi.org/10.1017/langcog.2022.8

Jones, R. H. (Ed.). (2015). *The Routledge handbook of language and creativity*. Routledge.

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170–210. https://doi.org/10.2307/411200

Kaufman, J. C., & Sternberg, R. J. (Eds.). (2019). *The Cambridge handbook of creativity* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781316979839

Kim, A., & Gilley, P. (2013). Neural mechanisms of rapid sensitivity to syntactic anomaly. *Frontiers in Psychology*, *4*.

Kuperberg, G. R., Choi, A., Cohn, N., Paczynski, M., & Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *Journal of Cognitive Neuroscience*, *22*(12), 2685–2701. https://doi.org/10.1162/jocn.2009.21333

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lauwers, P., & Willems, D. (2011). Coercion: Definition and challenges, current approaches, and new trends. *Linguistics*, *49*(6), 1219–1235. https://doi.org/10.1515/ling.2011.034

Lenth, R. (2023). *Emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.5*. https://CRAN.R-project.org/package=emmeans

Levin, B. (2018). Argument Structure. *Oxford Bibliographies in Linguistics*. https://doi.org/10.1093/obo/9780199772810-0099

Levin, B., & Rappaport Hovav, M. (2005). *Argument realization*. Cambridge University Press.

Libben, G., Derwing, B. L., & de Almeida, R. G. (1999). Ambiguous novel compounds and models of morphological parsing. *Brain and Language*, *68*(1), 378–386. https://doi.org/10.1006/brln.1999.2093

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, *52*(3), 436–459. https://doi.org/10.1016/j.jml.2005.01.013

Luka, B. J., & Choi, H. (2012). Dynamic grammar in adults: Incidental learning of natural syntactic structures extends over 48 h. *Journal of Memory and Language*, *66*(2), 345–360. https://doi.org/10.1016/j.jml.2011.11.001

Lukassek, J., Prysłopska, A., Hörnig, R., & Maienborn, C. (2017). The semantic processing of motion verbs: Coercion or underspecification? *Journal of Psycholinguistic Research*, *46*(4), 805–825. https://doi.org/10.1007/s10936-016-9466-7

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037/0033-295X.101.4.676

Manouilidou, C., & Almeida, R. G. de. (2013). Processing correlates of verb typologies: Investigating internal structure and argument realization. *Linguistics*, *51*(4), 767–792. https://doi.org/10.1515/ling-2013-0026

Mashal, N., & Faust, M. (2009). Conventionalisation of novel metaphors: A shift in hemispheric asymmetry. *Laterality*, *14*(6), 573–589. https://doi.org/10.1080/13576500902734645

Meßmer, J. A., Bader, R., & Mecklinger, A. (2021). The more you know: Schema-congruency supports associative encoding of novel compound words. Evidence from event-related potentials. *Brain and Cognition*, *155*, 105813. https://doi.org/10.1016/j.bandc.2021.105813

Michaelis, L. A. (2004). Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive Linguistics*, *15*(1), 1–67. https://doi.org/10.1515/cogl.2004.001

Michaelis, L. A. (2005). Entity and event coercion in a symbolic theory of syntax. In J.-O. Östman & M. Fried (Eds.), *Construction Grammars: Cognitive grounding and theoretical extensions* (pp. 45–88). John Benjamins.

Mitchell, D. C. (1987). Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 601–618). Erlbaum.

Müller, S., & Wechsler, S. (2014). Lexical approaches to argument structure. *Theoretical Linguistics*, *40*(1–2), 1–76. https://doi.org/10.1515/tl-2014-0001

Munat, J. (2015). Lexical creativity. In R. H. Jones (Ed.), *The Routledge handbook of language and creativity* (pp. 92–106). Routledge.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state

knowledge and animacy selection restrictions. *Journal of Memory and Language*, *67*(4), 426–448. https://doi.org/10.1016/j.jml.2012.07.003

Perek, F., & Hilpert, M. (2014). Constructional tolerance: Cross-linguistic differences in the acceptability of non-conventional uses of constructions: *Constructions and Frames*, *6*(2), 266–304. https://doi.org/10.1075/cf.6.2.06per

Piñango, M. M., Winnick, A., Ullah, R., & Zurif, E. (2006). Time-course of semantic composition: The case of aspectual coercion. *Journal of Psycholinguistic Research*, *35*(3), 233–244. https://doi.org/10.1007/s10936-006-9013-z

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT Press.

Pollatsek, A., Bertram, R., & Hyönä, J. (2011). Processing novel and lexicalised Finnish compound words. *Journal of Cognitive Psychology*, *23*(7), 795–810. https://doi.org/10.1080/20445911.2011.570257

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Pynte, J., Besson, M., Robichon, F.-H., & Poli, J. (1996). The time-course of metaphor comprehension: An event-related potential study. *Brain and Language*, *55*(3), 293–316. https://doi.org/10.1006/brln.1996.0107

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Rayner, K., & Sereno, S. C. (1994). Regressive eye movements and sentence parsing: On the use of regression-contingent analyses. *Memory & Cognition*, *22*(3), 281–285. https://doi.org/10.3758/BF03200855

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1290–1301. https://doi.org/10.1037/0278-7393.30.6.1290

Rösler, F., Pechmann, T., Streb, J., Röder, B., & Hennighausen, E. (1998). Parsing of sentences in a language with varying word order: Word-by-word variations of processing demands are revealed by event-related brain potentials. *Journal of Memory and Language*, *38*(2), 150–176. https://doi.org/10.1006/jmla.1997.2551

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Staub, A. (2007). The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(3), 550–569. https://doi.org/10.1037/0278-7393.33.3.550

Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., & Majewski, H. (2007). The time course of plausibility effects on eye movements in reading: Evidence from noun-noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(6), 1162–1169. https://doi.org/10.1037/0278-7393.33.6.1162

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2023). Rating norms should be calculated from cumulative link mixed effects models. *Behavior Research Methods*, *55*(5), 2175–2196. https://doi.org/10.3758/s13428-022-01814-7

Thorne, J. P. (1965). Stylistics and generative grammars. *Journal of Linguistics*, *1*(1), 49–59. https://doi.org/10.1017/S0022226700001018

Ungerer, T., Antal, C., & de Almeida, R. G. (2025, November 3). *Comprehension of valency coercion*. https://doi.org/10.17605/OSF.IO/UQKP4

Ungerer, T., & Hartmann, S. (2023). *Constructionist approaches: Past, present, future*. Cambridge University Press. https://doi.org/10.1017/9781009308717

van Gompel, R. P. G., & Pickering, M. J. (2001). Lexical guidance in sentence processing: A note on Adams, Clifton, and Mitchell (1998). *Psychonomic Bulletin & Review*, *8*(4), 851–857. https://doi.org/10.3758/BF03196228

van Trijp, R. (2015). Cognitive vs. generative construction grammar: The case of coercion and argument structure. *Cognitive Linguistics*, *26*(4), 613–632. https://doi.org/10.1515/cog-2014-0074

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews. Cognitive Science*, *4*(2), 125–134. https://doi.org/10.1002/wcs.1209

Vogel, R. (2023). Grammatical gaps, grammatical invention and grammatical theory. In T. Strobel & H. Weiß (Eds.), *Grammatical gaps: Definition, typology and theory* (pp. 15–50). Buske.

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, *14*(4), 770–775. https://doi.org/10.3758/bf03196835

Warren, T., McConnell, K., & Rayner, K. (2008). Effects of context on eye movements when reading about possible and impossible events. *Journal of Experimental Psychology.*

Learning, Memory, and Cognition, *34*(4), 1001–1010. https://doi.org/10.1037/0278-7393.34.4.1001

Warren, T., Milburn, E., Patson, N. D., & and Dickey, M. W. (2015). Comprehending the impossible: What role do selectional restriction violations play? *Language, Cognition and Neuroscience*, *30*(8), 932–939. https://doi.org/10.1080/23273798.2015.1047458

Williams, A. (2015). *Arguments in syntax and semantics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139042864

Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, *41*(2), 105–128. https://doi.org/10.1007/s10936-011-9179-x

Yoon, S. (2016). Gradable nature of semantic compatibility and coercion: A usage-based approach. *Linguistic Research*, *33*(1), 95–134. https://doi.org/10.17250/khisli.33.1.201603.005

Yoon, S. (2019). Coercion and language change: A usage-based approach. *Linguistic Research*, *36*(1), 111–139. https://doi.org/10.17250/khisli.36.1.201903.005

Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832

Appendix A

List of Experimental Materials

No.	Text passage	Acceptability	Creativity	Plausibility
01	Frank swallowed a red chili pepper at the dinner table. Tears streamed from his eyes, and he reached blindly for his napkin. Unable to control himself,			
	(P) Frank pushed his napkin off the table	5	1	5
	(C) Frank sneezed his napkin off the table	2	3	2
	(A) Frank arrived his napkin off the table	1	4	1
	… and knocked over a few of the wine glasses.			
02	Sharon was arguing with her husband in the kitchen. They raised their voices as the discussion grew more and more heated. In the end,			
	(P) Sharon shoved her husband out of the kitchen	5	2	5
	(C) Sharon yelled her husband out of the kitchen	4	3	4
	(A) Sharon relied her husband out of the kitchen	1	3	2
	… and slammed the door with a loud bang.			
03	Linda had filed for divorce when her husband's lawyers suddenly came to her house. They offered her money if she changed her mind. Keeping her cool,			
	(P) Linda threw the lawyers out of her house	5	2	5
	(C) Linda shrugged the lawyers out of her house	3	4	3
	(A) Linda wandered the lawyers out of her house	3	3	2
	… and felt proud for not giving in to the pressure.			
04	James was boxing against an amateur. His opponent was so scared that he hardly managed to climb into the ring. In the first round,			
	(P) James knocked his opponent out of the ring	5	3	5
	(C) James stared his opponent out of the ring	4	4	4
	(A) James limped his opponent out of the ring	2	4	2
	… and ended the fight before it had even begun.			
05	John and his dance partner entered the ballroom. They immediately spotted a rich buffet at the far end of the hall. Without hesitation,			
	(P) John guided his partner across the ballroom	5	2	5
	(C) John tangoed his partner across the ballroom	3	2	4
	(A) John emerged his partner across the ballroom	2	3	3
	… and indulged in some of the sweets and desserts.			
06	Mary found an empty table in the jazz bar. Seeing a good-looking waiter, she tried to catch his attention. When their eyes met,			
	(P) Mary called the waiter over to her table	5	2	5
	(C) Mary winked the waiter over to her table	4	4	4
	(A) Mary knelt the waiter over to her table	2	3	2
	… and ordered her favorite cocktail off the menu.			

07	Patrick and his friend Max stayed at a hostel dormitory. Before going to bed, Max teased Patrick with some mean jokes. In return,			
	(P) Patrick kicked his friend out of the dormitory	5	1	5
	(C) Patrick snored his friend out of the dormitory	3	4	3
	(A) Patrick agreed his friend out of the dormitory	1	3	1
	… and had the room for himself for the rest of the night.			
08	Romeo was waiting in the street below his lover's balcony. As the girl appeared, he strummed the first chords on his guitar. With his music,			
	(P) Romeo lured his lover down to the street	5	4	4
	(C) Romeo sang his lover down to the street	3	4	3
	(A) Romeo swam his lover down to the street	1	4	1
	… and made a passionate speech about his feelings.			
09	Betty saw the neighbor's children climb over the fence and into her garden. They stepped onto her favorite flower bed. Waving her hands angrily,			
	(P) Betty hustled the children back over the fence	4.5	3	4.5
	(C) Betty frowned the children back over the fence	3	4	2
	(A) Betty panicked the children back over the fence	2	4	3
	… and complained to the parents later that evening.			
10	Anne was walking her poodle when a jogger passed them. The dog barked and jumped excitedly around the man. Feeling a bit embarrassed,			
	(P) Anne dragged her poodle away from the jogger	5	2	4
	(C) Anne whistled her poodle away from the jogger	3	3	3
	(A) Anne squatted her poodle away from the jogger	1	4	2
	… and continued down the path in the other direction.			
11	Susan and her children went for a walk along the sea cliff. The boys were playing dangerously close to the edge. Growing increasingly anxious,			
	(P) Susan pulled her children away from the cliff	5	2	5
	(C) Susan gestured her children away from the cliff	5	2	5
	(A) Susan belonged her children away from the cliff	1	2	1
	… and told herself to be more careful the next time.			
12	Jason found an old book in his grandfather's library. As he touched the cover, some dust came off the surface. Feeling a tickle in his throat,			
	(P) Jason brushed the dust off the book	5	2	5
	(C) Jason coughed the dust off the book	3	4	3
	(A) Jason existed the dust off the book	1	4	1
	… and discovered some beautiful illustrations underneath.			
13	Nancy heard the door chime as some customers entered her shop. They were pretty drunk and tried to flirt with her. In a polite but firm manner,			
	(P) Nancy ushered the customers out the door	5	2	5
	(C) Nancy smiled the customers out the door	3	2	3
	(A) Nancy leaped the customers out the door	2	4	2
	… and returned to her tasks with a sigh of relief.			

14 The audience saw the singer waiting next to the concert stage. Finally, the previous performer's turn was over. With enthusiastic cheers,

(P) the audience called the singer onto the stage	5	4	5
(C) the audience clapped the singer onto the stage	2	3	4
(A) the audience napped the singer onto the stage	1	3	2

… and listened reverently to the first chords.

15 The boss met the employee in his office. The assistant presented new business ideas, but the boss did not like any of them. Losing his patience,

(P) the boss kicked the employee out of his office	5	2	5
(C) the boss yawned the employee out of his office	2	4	2
(A) the boss dozed the employee out of his office	2	4	2

… and continued with other items on the agenda.

16 The generals hated the king ever since he had ascended the throne. They met in secret and planned to overthrow him. Using threats and intrigue,

(P) the generals forced the king off the throne	5	2	4
(C) the generals conspired the king off the throne	4	3	4
(A) the generals consisted the king off the throne	1	3	2

… and seized power in the country for a long time.

17 The janitor was about to close the gates of the garden. Just at that moment, some last visitors arrived at the exit. With an impatient smile,

(P) the janitor ushered the visitors through the gates	5	2	5
(C) the janitor nodded the visitors through the gates	3	2	4
(A) the janitor reacted the visitors through the gates	1	2	1

… and finished his shift soon afterwards.

18 The judge read out his verdict in the courtroom. The defendant cried out and started swearing violently. Furious at this interruption,

(P) the judge ordered the defendant out of the courtroom	5	2	5
(C) the judge roared the defendant out of the courtroom	3	4	3
(A) the judge leaned the defendant out of the courtroom	2	4	3

… and declared the trial to be over.

19 The managers disliked one of the colleagues in their project. They spread false stories to get rid of the person. Using this strategy,

(P) the managers forced their colleague out of the project	5	2	4
(C) the managers rumored their colleague out of the project	2	3	2
(A) the managers depended their colleague out of the project	1	3	1

… and hired a new staff member instead.

20 A policeman was regulating the traffic at the intersection. Some teenagers were jeering at him while waiting to cross. In a grumpy tone,

(P) the policeman directed the teenagers across the intersection	5	2	5
(C) the policeman barked the teenagers across the intersection	2	3	2
(A) the policeman prayed the teenagers across the intersection	2	2	2

… and turned his attention back to the cars.

21	The principal stood in front of the school building. Suddenly, he saw two pupils sneak out onto the street. In a fit of rage,			
	(P) the principal hustled the pupils back into the building	4	2	4
	(C) the principal screamed the pupils back into the building	4	4	3
	(A) the principal sprinted the pupils back into the building	3	4	3
	… and gave the culprits two hours of detention.			
22	The queen had been lying sick in bed for several weeks. Growing weaker and weaker, she sent for her best doctors. Once everyone had arrived,			
	(P) the queen called the doctors to her bedside	5	2	4
	(C) the queen sighed the doctors to her bedside	2	3	3
	(A) the queen paused the doctors to her bedside	2	3	2
	… and asked if any cure might be found.			
23	The students were pranking their teacher. When the man entered the classroom, a bucket of paint poured down on his head. Showing no pity,			
	(P) the students pushed the teacher out of the classroom	5	3	4
	(C) the students laughed the teacher out of the classroom	4	2	4
	(A) the students crawled the teacher out of the classroom	1	3	2
	… and congratulated each other on their success.			
24	The hockey players cried out as their teammate got injured. He received medical treatment outside the ice rink. When things looked better,			
	(P) the players pulled their teammate back into the rink	5	2	5
	(C) the players cheered their teammate back into the rink	5	3	5
	(A) the players glanced their teammate back into the rink	2	4	2
	… and doubled their efforts to win the match.			

Note. Text passages with context sentences and three target versions: prototypical (P), coerced (C), and anomalous (A). Columns on the right show norming ratings ($N = 21$) for the median acceptability, creativity, and plausibility of each target (1 = lowest, 5 = highest).

Appendix B

Argument-Linking Profiles of Coerced Verbs

Verb	Transitive uses with animate NP in COCA (%)	Transitive uses with inanimate NP in COCA (%)	Intransitive and other uses in COCA (%)	Expected type of NP violation
barked	0	13	87	animacy
cheered	26	9	65	animacy
clapped	7	48	45	animacy
conspired	0	0	100	transitivity
coughed	0	25	75	animacy
frowned	0	0	100	transitivity
gestured	2	1	97	transitivity
laughed	0	1	99	transitivity
nodded	0	3	97	transitivity
roared	0	3	97	transitivity
rumored	28	21	51	animacy
sang	3	41	56	animacy
screamed	0	8	92	animacy
shrugged	2	13	85	animacy
sighed	0	1	99	transitivity
smiled	0	2	98	transitivity
sneezed	0	3	97	transitivity
snored	0	0	100	transitivity
stared	3	0	97	transitivity
*tangoed**	0	0	100	transitivity
whistled	1	10	89	animacy
winked	0	2	98	transitivity
yawned	0	0	100	transitivity
yelled	1	8	91	animacy

Note. Corpus-attested argument-linking patterns of verbs in the coerced condition, based on manual annotation of 100 instances of each verb (only 14 instances in the case of *tangoed*) from the *Corpus of Contemporary American English* (COCA; Davies, 2008). Verbs with fewer than 5% of transitive attestations are classified as giving rise to transitivity violations; all other verbs are assumed to produce animacy violations.