A Construction Grammar Corpus of Varying Schematicity: A Dataset for the Evaluation of Abstractions in Language Models

Claire Bonial*, **Harish Tayyar Madabushi**†
*DEVCOM U.S. Army Research Lab, USA
†University of Bath, UK
claire.n.bonial.civ@army.mil, htm43@bath.ac.uk

Abstract

Large Language Models (LLMs) have been developed without a theoretical framework, yet we posit that evaluating and improving LLMs will benefit from the development of theoretical frameworks that enable comparison of the structures of human language and the model of language built up by LLMs through the processing of text. In service of this goal, we develop the Construction Grammar Schematicity ("CoGS") corpus of 10 distinct English constructions, where the constructions vary with respect to schematicity, or in other words the level to which constructional slots require specific, fixed lexical items, or can be filled with a variety of elements that fulfill a particular semantic role of the slot. Our corpus constructions are carefully curated to range from substantive, frozen constructions (e.g., LET-ALONE) to entirely schematic constructions (e.g., RESULTATIVE). The corpus was collected to allow us to probe LLMs for constructional information at varying levels of abstraction. We present our own probing experiments using this corpus, which clearly demonstrate that even the largest LLMs are limited to more substantive constructions and do not exhibit recognition of the similarity of purely schematic constructions. We publicly release our dataset, prompts, and associated model responses.

Keywords: Construction Grammars, CxGs and NLP, Large Language Models

1. Introduction

Large Language Models (LLMs) offer unprecedented opportunities for the advancement of both NLP and AI by supporting natural language interaction with a variety of different computational systems, but also by providing a store of cultural and commonsense knowledge. However, before taking advantage of LLMs in applications that require consistent, trustworthy, transparent behavior (e.g., medical and disaster-relief applications), the limits of this linguistic and commonsense knowledge must be better understood. The current generation of LLMs are not only opaque in terms of their functioning but also pushing the boundaries of what is achievable in terms of their size. This, along with recent findings suggesting that these models may struggle with reasoning (Lu et al., 2023), highlights the necessity of assessing the cognitive foundation of LLM performance. Understanding the limitations and facilitating targeted improvements will be aided by a theoretical framework that allows us to compare the structures of human grammar to the model of human language built by LLMs.

The bottom-up, usage-based Construction Grammar (CxG) model of language provides an ideal foundation for this comparison, as it may have better comparative and explanatory capacity in the context of modeling the bottom-up, processing-based linguistic capabilities of LLMs (Goldberg, to appear). The knowledge of LLMs, both linguistic and otherwise, is built up through training on text alone. While this may be considered similar to a "usage"-based model of human processing, it is also in contrast to the human experience where language is embodied, and each token of linguistic experience is enriched by cross-modal association (Bybee, 2006). Thus, the CxG framework is ripe for comparing the processes of bottom-up processing of language and subsequent abstraction over the individual experiences of language that result in higher-order grammatical structure.

In addition, recent evidence suggests that LLMs are particularly good at capturing formal linguistic competence or information pertaining to language rules as opposed to functional linguistic competence (Mahowald et al., 2023). Functional linguistic competence includes the capabilities required for real-world comprehension and use. Consequently, examining the concept of schematicity in constructions provides an ideal means of evaluating these models. We believe that this method provides an opportunity for addressing the missing components in LLMs, making them more powerful and potentially sample efficient.

In this research,[1] we develop a novel collection of over 500 corpus instances of 10 unique English constructions, carefully curated to test varying levels schematicity—from purely substantive constructions that are fixed in their form (e.g., LET-ALONE) to purely schematic argument structure constructions

[1] https://github.com/H-TayyarMadabushi/Construction_Grammar_Schematicity_Corpus-CoGS

LREC-COLING 2024, pages 243–255
20-25 May, 2024.

(e.g., RESULTATIVE). We provide a listing of over 500 corpus instances from the CoGS corpus and make our evaluation results, including model generated response, available for public release. The collection of this corpus is not at all trivial, and we describe the challenges of detecting usages in corpora that are not currently annotated syntactically or semantically in line with the tenets of CxG.

In our experiments, we employ some of the most advanced LLMs currently available, with a minimum of 175 billion parameters and trained on nearly the entirety of web data. Our results show that, although these LLMs can effectively distinguish instances of less schematic constructions, their capacity to do so significantly declines as the schematicity of constructions increases. This observation remains consistent across the LLMs we assess and with the various prompts we employ in the tests, ensuring its generalizability. Table 1 presents an overview of our results, with further details in Section 5.2.

Abstraction Level	**GPT-3.5 (%)**	**GPT-4 (%)**
Purely Substantive	84.00	98.34
Partially Schematic	75.17	92.67
Fully Schematic	54.00	62.33

Table 1: Results of our most effective probe. Consistently, an increase in schematicity results in lower performance (see §5.2)

In the sections to follow, we provide a theoretical background of CxG (§2) and summary of related work (§3), which motivates our collection of constructions organized according to schematicity (§4). With this novel corpus, we probe LLMs for constructions at varying levels of schematicity (§5). We present our results which demonstrate clear and consistent LLM performance differences across substantive, partially substantive, and purely schematic constructions (§5.2). These results provide fodder for our conclusions and paths for our future work introducing additional resources to support synergistic research advancing LLMs with the CxG theoretical framework (§6).

2. Theoretical Background

Distinct branches of Construction Grammar (CxG) (see Hoffmann and Trousdale (2013) for a comprehensive description) agree upon the notion that all aspects of language rest upon a central cognitive unit: the construction. Constructions are form-meaning pairings, in many ways similar to the original Saussurean sign—the pairing of a *concept* ("signified") and a sound image ("signifier") (De Saussure et al., 1916). The notion of the construction as the central unit of language is in contrast to the foundational generative assumption that the lexicon (where word meanings are stored) and syntax (where the rules of how to put together those words are specified) are separate modules of human linguistic knowledge (e.g., Chomsky (2014)). In CxG, not only are word forms and their meanings constructions, but also morphemes and certain combinations of words or phrases.

Most theories of CxG also posit that all of these constructions make up a speaker's mental grammatical knowledge, termed a 'constructicon,' as opposed to a lexicon.[2] If all constructions are stored from morphemes up through phrases of the grammar, then there is no need for speakers to also store and access knowledge of the syntactic rules guiding the combinatoric potential of the language.

The constructicon is characterized by a taxonomic organization with inheritance potential between related constructions. This structure arises from the usage-based character of CxG, wherein the constructions that are commonly heard by a language speaker are initially stored as unanalyzed whole holophrases, for example, *want milk* (Tomasello, 2006). With continued exposure to the language, leveraging domain-general cognitive processes such as pattern-finding, speakers recognize commonalities across constructions and begin to generalize both fixed and open slots of constructions; for example *want juice, want mommy*. Throughout this process, there is abstraction over the slots of a construction such that grammatical knowledge approximating parts of speech arises (Tomasello, 2005). As the constructicon is built up with each token of experience of the language, the emerging taxonimical structure is one wherein the lowest levels of the taxonomy are the actual constructs experienced in language (Hoffmann, 2022). As generalization and abstraction occur over these constructs, the higher-order parent constructions are made up of increasingly abstract, schematic slots and decreasing numbers of substantive, fixed word forms.

While there is some agreement in CxG that the constructicon must have taxonomic structure, it is not clear exactly what constructions exist in a constructicon beyond the constructs that we know a speaker has been exposed to. However, evidence for generalization to the level of purely schematic constructions comes in the form of Argument Structure Constructions (Goldberg, 1995). Argument Structure Constructions are purely schematic—that is, there is no substantive word or phrase that is required for speakers to recognize and generate constructs of this construction. Instead, speakers are able to recognize the combination of particular slots as carrying a particular meaning, even if that particular construct hasn't been heard before. For

[2] Theories differ as to whether all constructions a speaker is exposed to are stored (e.g., Bybee (2010)), or only those that cannot be decomposed into the combination of other stored constructions (e.g., Jackendoff (2010)).

example, *The boy XXXX the YYYY into the ZZZZ*—even if we do not know or understand several of the content words of this utterance, a native speaker can glean the event is likely one of caused motion given the nature of the arguments. Mainstream generative grammar has traditionally posited that the semantic roles of the participants of a particular event are assigned by the verb, which denotes and projects the semantics of that event (e.g., Jackendoff (1992)). In contrast, CxG posits that the semantics stem from the construction itself, which explains why speakers recognize the unique semantics of this construction even when instantiated in novel and creative ways.

It is against this theoretical background that we begin to consider and probe the linguistic information that LLMs have access to. In probing LLMs for constructional knowledge, one must take into account the fact that there is no consensus as to what the constructions of a language are, and, in particular, how abstract or schematic the constructions of the constructicon are at the highest levels of the taxonimic structure. Thus, if probing studies show that LLMs do or do not have knowledge of a particular construction that is not agreed upon in CxG, then we may also bring to light valuable areas for future theoretical and psycholinguistic research.

3. Related Work in CxG and NLP

Distributional word properties captured by LLMs can be powerful in a variety of different NLP applications, and have been leveraged in CxG approaches to characterize the words that are compatible with particular slots of a construction, including the Dutch Causative (Levshina and Heylen, 2014) and the verbal slots in English V the hell out of constructions and Way-manner constructions (Perek, 2018). Rambelli et al. (2019) propose a usage-based framework for representing constructions via vectors, but conclude with a note that in future work the distributional-semantic vectors should be paired with an accurate formalization of the internal structure of the construction.

Dunn (2017) describes data-driven induction of a constructicon wherein distributional vector semantics provide the "meaning" pole of the form-meaning pairings, and his CxG induction algorithm arrives at the form pole. His procedure results in constructional candidates which Dunn terms 'first- order constructions' (Dunn, 2023, 9), such as [VERB – ADP – DET – <521>], including usages like *come to this house* and *lying on the floor*. The '<' bracketed number in the construction indicates a cluster of distributionally similar words, here, locations such as *house* and *floor*. Dunn acknowledges that the data-driven constructions resulting from his process do not necessarily jibe with meaningful constructions posited in the theories of CxG introduced here, which have experimental evidence of their feasibility within the mental constructicon (e.g., Kaschak and Glenberg (2000); Johnson and Goldberg (2013)). Dunn distinguishes such first-order constructions from second-order constructions, which include first-order (nominal and verbal) constructions as the fillers of particular slots within the second-order construction, and are therefore characterized by a greater level of schematicity. Dunn outlines methodologies for detecting and including second-order constructions in the same data-driven approach, but leaves this for future work. Thus, Dunn's data-driven approach is limited to more substantive constructions at the lowest levels of the constructicon.

Tayyar Madabushi et al. (2020) directly probe a variety of BERT-based models for access to knowledge of several of the constructions proposed in Dunn (2017). This research prompts BERT to distinguish between sentences that instantiate a particular construction and those that do not. The models evaluated included a base BERT model compared with BERT clones that have been infused with, or pre-trained on, constructional information in the form of training sentences with the construction explicitly identified. The authors altered the frequencies of the constructions used for pre-training, such that certain models were trained on low-frequency constructions, while others were trained on high-frequency constructions. The authors show that even when the clone model is pre-trained on low frequency constructions, which one would hypothesize are less likely to already be present in the model, that the pre-trained clones with constructional information do not perform significantly better than the base BERT model on a downstream task probing constructional information. Thus, the authors conclude that BERT already has access to such constructional information. However, a limitation of this finding acknowledged by the authors is that the constructions leveraged for this study were those stemming from the data-driven approach of Dunn (2017), so the constructions successfully probed for are from the lowest levels of the constructional hierarchy, representing more substantive constructions (Tayyar Madabushi et al., 2024).

In contrast, Li et al. (2022) probe models of varying sizes for access to knowledge of purely schematic argument structure constructions, including Ditransitive, Resultative, Caused-motion, and Removal constructions. The authors recreate psycholinguistic studies, adapted for application to language models. Here, we focus on the first study: a sentence sorting task, in which speakers/models are prompted to sort sentences by similarity. The sentences are carefully curated, such that the constructions of interest are instantiated with a variety of different lexical verbs, and the lexical verbs (e.g., *cut, slice*) instantiating distinct constructions stem from semantically similar verb classes. Thus, the study explores whether speakers sort the sentences according to lexical verbal semantics (as would be predicted by mainstream generative grammar approaches) or constructional semantics. The authors find that while the smallest language model with 1 million parameters, MiniBERTas (Warstadt et al., 2020), groups the sentences according to lexical semantics, the largest model with 30 billion parameters, RoBERTa (Liu et al., 2021), groups sentences according to constructional semantics. This result mirrors the psycholinguistic studies, which demonstrated that language learners group sentences according to lexical semantics, while native speakers group sentences according to constructional semantics.

Bunzeck and Zarrieß (2023) aim to discover how fast constructional entrenchment and generalizations arise, given differences in model size. Specifically, the authors use RoBERTa-based models of various sizes in a perturbed masking approach to study the influence of several factors on sentence embeddings: token position, token length (in characters), sequence length (number of tokens), and construction type (IMPERATIVE, TRANSITIVE, WH-QUESTION constructions). The authors find that across all model sizes, the token position has the most influence, with tokens at the end of a sequence having the greatest influence on the sequence embedding. The effect of construction type is somewhat positive, but varies across parts of speech and across the models. Nouns, proper nouns, and punctuation hold the most influence across all constructions; however, for larger models, verbs begin to gain some additional influence. Overall, the authors conclude that the construction types do not hold much explanatory value in understanding the role of entrenchment in LLMs, and that not all models learn the same structures.

Weissweiler et al. (2022) add to this picture with their analysis of the COMPARATIVE-CORRELATIVE (e.g., *The higher you fly, the harder you fall*). The authors probe how well LLMs can capture both the syntactic and semantic information of the construction. To determine the LLM's ability to recognize the syntactic, form side of the construction, the authors test the models for their ability to identify instances of the construction in corpus examples as well as synthetic data. The authors find that a variety of different BERT-based LLMs are able to identify and distinguish cases of the COMPARATIVE-CORRELATIVE (which we would consider a partially substantive construction). To evaluate the ability of the LLMs to interpret the semantics of the COMPARATIVE-CORRELATIVE, the authors leverage a downstream task that requires correct interpretation of the construction. The models perform poorly, close to chance, on this semantic evaluation. Thus, this study shows that while BERT-based models have access to syntactic information about the COMPARATIVE-CORRELATIVE, they are unable to interpret the meaning of the construction.

Veenboer and Bloem (2023) contrast collostructional analysis (Stefanowitsch and Gries, 2003) results to output from BERT on two tasks: First, the prediction of words in the schematic slots of constructions; second, the examination of the output embeddings to determine whether less lexicalized constructions are clustered together in semantic space. The authors focus on two constructions that already have collustructional analysis results: X WAITING-TO-HAPPEN and the DITRANSITIVE. We would characterize the former as more substantive, whereas the latter is a purely schematic construction. Regarding the X WAITING-TO-HAPPEN construction, the authors observe that BERT's output candidates for X capture the semantics of the slot—events that are imminent—and all of the output candidates are felicitous, although the top-ranked BERT output, *event*, was exhibited with low frequency. For the DITRANSITIVE construction, both collostructional analysis and BERT results output *give* as the most likely candidate to sit in the verb slot. The inspection of the output embeddings showed that BERT places the DITRANSITIVE and the near-synonymous dative variants clustered together, indicating that the embedding space mirrors the meaning of the construction. The authors conclude that they find no evidence that BERT clusters by constructional form when meaning is similar—the model does not fully distinguish constructions with similar meaning but different form. The authors also conclude that BERT represents more lexicalized, substantive constructions better than abstract constructions.

4. Corpus Collection

Related work leads us to hypothesize that LLMs may have access to the lower levels of a language's constructional hierarchy—i.e., the constructions that are more substantive and the constructs of a construction that are attested frequently. These frequently attested tokens of a construction are filled with lexeme constructions that are strongly entrenched in that construction (for example, "make" within the RESULTATIVE construction, and "give" within the DITRANSITIVE). However, we hypothesize that LLMs may not have access to the higher levels of a language's constructional hierarchy—i.e., the constructions that are purely schematic. LLMs may not be able to distinguish between two fully schematic constructions where the form pole is very similar, but the meaning pole (i.e., semantic roles associated with particular constructional slots) is different. For example, distinguishing between LABELING constructions, such as *I consider him a friend*, and the DITRANSITIVE: *I called him a cab.*

Substantive	Let-alone
	Much-less
Partially Substantive	Way-manner
	Comparative-correlative
	Conative
	Causative-with
Schematic	Caused-motion
	Intransitive-motion
	Ditransitive
	Resultative

Table 2: Ten constructions of focus, organized from most substantive to purely schematic.

To test this hypothesis, we curated a dataset of ten distinct constructions that range from substantive with frozen phrases to fully schematic constructions. We collected a minimum of 50 corpus instances of each construction, for a total corpus size of 524 constructs, from several different corpora: The FrameNet Constructicon (Fillmore et al., 2012),The Corpus of Contemporary American English (COCA) (Davies, 2008), and the Abstract Meaning Representation 3.0 release corpus from the Linguistic Data Consortium (LDC2020T02) (Knight et al., 2020). We also included a handful of positive examples of argument structure constructions from linguistic research articles. The constructions are listed in Table 2 from substantive to schematic (see Table 4 in Appendix A for a fuller summary with examples). In this section, we motivate our choice of constructions and describe how we defined and collected instances of the construction.

4.1. Construction Selection

Let-alone, Much-less

We select two related, substantive constructions, LET-ALONE and MUCH-LESS. The LET-ALONE construction is well described in Fillmore et al. (1988), from which we draw our outline of the semantics of the construction: LET-ALONE coordinates two elements before and after the contiguous, fixed phrase *let alone*. The two elements represent points on a scale of some property, which is largely implicit and understood only from the background and cultural context. For example, *I doubt he made colonel in World War II, let alone general* invokes some cultural knowledge of the military ranks. The element that follows LET-ALONE is posed with greater illocutionary force and is "higher" on the implicit scale—If I doubt that he made colonel, which is a lesser rank, then I doubt all the more that he made general. Notably, a speaker's familiarity with the construction can overcome a gap in cultural knowledge, such that one can glean the points on the implicit scale of each element without necessarily having knowledge of that scale. Fillmore et al. (1988) also describe the nearly identical syntactic and semantic behavior of MUCH-LESS, exemplified by *Haven't even compiled it, much less tested it.*

Way-manner

We also selected what we term the WAY-MANNER construction as a partially substantive construction. This construction is described and included in the FrameNet Constructicon, which distinguishes three related constructions: WAY-MANNER (e.g., *He whistled his way through the crowd*), WAY-MEANS (e.g., *He elbowed his way through the crowd*), and WAY-NEUTRAL (e.g., *He made his way through the crowd*). The WAY-MANNER and WAY-MEANS constructions are described syntactically as a verb exceptionally taking "one's way" as the object, where the possessive pronoun is coreferential with the subject of the verb. The object is followed by an obligatory prepositional phrase (PP) specifying a path. The semantics are described as that of motion of the verbal subject towards or along the path specified by the PP, characterized by the particular manner or means specified by the lexical semantics of the verb (e.g., whistling, elbowing).

Comparative-correlative

We collected instances of what we term the COMPARATIVE-CORRELATIVE, also sometimes illustratively named "The X-er, The Y-er" as another partially substantive construction. This construction has been treated in theoretical linguistic research (Goldberg, 2003) and NLP research, including both efforts to annotate the constructional semantics using Abstract Meaning Representation (AMR) (Banarescu et al., 2013; Bonial et al., 2018), as well as a probing study (Weissweiler et al., 2022). We follow the syntax and semantics outlined in Goldberg (2003) and Weissweiler et al. (2022), which characterize the COMPARATIVE-CORRELATIVE as coordinating two elements, both of which begin with "the" and are followed by a comparative adjective or adverb. Semantically, the two elements are in a causal or temporal relation, where the second element is dependent upon the first element. For example, *The longer this goes on, the worse his odds get* can be rephrased with a conditional: *If this goes on longer, then his odds will get worse.*

Conative, Causative-with

We selected two constructions that are more schematic, barring a single fixed word that is required in a particular slot of the construction: The CONATIVE and CAUSATIVE-WITH. We leverage semantic and syntactic descriptions from Hoffmann (2022), who calls both constructions 'quasi-argument structure constructions,' due to their almost entirely schematic slots. The CONATIVE is characterized by a prepositional object of the verb that requires the fixed word "at." Semantically, the construction denotes effort to accomplish the action of the verb with the prepositional argument serving as the target of that action. However, the construction does not necessarily entail success, and as a result is also compatible with iterative actions. For example, *She kicked at the ball (over and over again).* Causative-with also requires an oblique argument that begins with the fixed word "with." Semantically, the construction conveys an affected patient mapping to the direct object of the verb, where this patient is also the ground upon which the theme argument mapping to the "with" oblique is applied. For example, *They heaped their plates with food.*

Caused-motion, Intransitive-motion, Ditransitive, Resultative

The remaining four constructions of focus are all fully schematic Argument Structure Constructions described in detail in Goldberg (1995): Caused-motion, Intransitive-motion, Ditransitive, and Resultative. The Caused-motion construction takes the syntactic shape of NP V NP PP, where the subject is a causer, the object is the theme in motion, the PP specifies the path, destination, or source, and the V specifies the action done by the causer subject that results in motion. For example, *They laughed the actor off the stage.* The Intransitive-motion construction takes the syntactic shape of NP V PP, where the subject is a theme in motion, the PP specifies the path, destination, or source, and the V specifies the action done by the theme subject that accompanies or causes the motion. For example, *The fly buzzed into the room*. The Ditransitive takes the syntactic shape of NP V NP NP, where the first postverbal NP is a recipient and the second NP is the theme being transfered to this recipient; the V specifies the the action of the causer subject that facilitates or characterizes the manner of the transfer. We follow Goldberg's definition that semantically the construction requires the "intent" to transfer a concrete or abstract (often informational) entity, but not necessarily a completed, successful transfer. For example, *She kicked me the ball (but I wasn't able to intercept it).* The Resultative takes the syntactic shape of NP V NP OBL, where the postverbal NP is a patient undergoing a change of state, the OBL can be a variety of different phrasal types including ADJ and PP, and the V specifies the action of the causer subject that facilitates the change. We follow Goldberg's semantic requirement that there is a change of state in the patient. For example, *Firefighters cut the man free* (the man becomes free). This rules out false positives with similar syntactic shape, such as *They kept the bird alive,* which do not meet this semantic requirement (the bird doesn't become alive, but simply remains alive).

4.2. Finding Corpus Usages

In this section, we describe our methodology for searching corpora for the ten constructions of interest. Notably, it becomes increasingly difficult to pinpoint instances of the more schematic constructions. There are few of the valuable constructicons like the FrameNet Constructicon, and this is far from comprehensive. Thus, for constructions not included in the FrameNet Constructicon, we conducted automatic searches of semantic-role-labeled AMR data, and where this was not available, we conducted automatic searches of the Penn Treebank tagged COCA corpus. As a result, we were largely relying on semantic and syntactic information stemming from a lexical-semantic, phrase-structure syntactic grammatical framework in order to attempt to detect our constructions, which cross-cut a variety of lexical and syntactic categories.

The precise collection procedure differed depending upon the features of the source corpus and the target construction; details are provided for each construction below. Given that there is no single corpus explicitly marking these constructions, leveraging unique collection procedures for each construction facilitated the most efficient search and targeting of the construction. The instances were recorded in a shared spreadsheet where each instance was inspected by two linguists and native English speakers, trained in CxG and semantic role labeling resources, including AMR. If the linguists did not agree that a given instance was a positive example of the construction with a meaning that was clearly understandable to native speakers, then the instance was thrown out. This process continued until there were at least 50 instances of each construction.

Notably, we filtered out cases of the fully schematic constructions instantiated by the most frequent, canonical verbs of that construction. Specifically, we filtered out instances of *give* in the Ditransitive, *make* in the Resultative, and instances of any verbs with motion lexical semantics from the Intransitive and Caused-Motion constructions. We acknowledge that these are all clear cases of the construction; however, in these cases, the lexical verb semantics and constructional semantics perfectly overlap. Thus, these cases do not probe constructional knowledge of LLMs, as this could instead be attributed to knowledge of frequently attested and entrenched lexical verb patterns.[3]

The Let-alone and Much-less constructions do not differ with respect to semantics necessarily, but they do differ in their realizations, and therefore differed in how we were able to collect instances of the construction. In both cases, we merely searched the constructional terms, requiring that they be contiguous. However, while the

[3]While this may be viewed as evidence that LLMs have a more lexical, compositional and Generative (in the sense of (Chomsky, 2014)) knowledge of language, such theories do not provide any explanatory mechanism enabling a better understanding or prediction/diagnosis of which verbal patterns are recognized and which are not.

search for LET-ALONE returned no false positives, the collocation “much” + “less” is found in a variety of quantity comparisons, so these false positives had to be removed. Thus, the collocation “let”+“alone” is highly entrenched as the LET-ALONE construction, whereas “much”+“less” may be realized in multiple distinct constructions. We collected corpus instances listed for LET-ALONE in the FrameNet Constructicon (Fillmore et al., 2012). We collected MUCH-LESS instances through search and manual filtering (leveraging the syntactic and semantic definition described above) of the COCA corpus (Davies, 2008).

Our WAY-MANNER constructs were collected from the FrameNet Constructicon. Specifically, we conflated and collected the corpus instances listed in the FrameNet Constructicon for both WAY-MANNER and WAY-MEANS.

Because the AMR corpus contains an explicit constructional “roleset” for the COMPARATIVE-CORRELATIVE, we collected most of our instances from this corpus by simply searching for the `correlate-91` roleset, which is devoted to exclusively marking up the semantic arguments of this construction (Bonial et al., 2018). The positive corpus examples included in Weissweiler et al. (2022) were also collected into our examples.

To collect corpus instances of the CONATIVE and CAUSATIVE-WITH constructions, we specified the argument structure of the construction and the fixed word and searched for these patterns in COCA (NP V PP-at NP for the CONATIVE and NP V NP PP-with for CAUSATIVE-WITH). This process results in a somewhat overwhelming number of false positives, but these were manually filtered until we obtained the desired number of instances.

We collected corpus instances of the CAUSED-MOTION or INTRANSITIVE-MOTION constructions by searching the AMR corpus for semantic roles compatible with the PPs of these constructions: `destination`, `source`, and `path`. AMR leverages the PropBank (Palmer et al., 2005) lexicon of “rolesets” which specify a listing of the semantic arguments of a given (generally lexical) relation in the form of numbered “Args” (Arg0, Arg1, Arg2...). The PropBank/AMR lexicon uses numbered arguments for frequent or “core” arguments of a given relation. There are also a variety of modifier or “adjunct” arguments that can be found with a variety of different relations. As a result of this setup, a search for the cases of modifier arguments that are specified as `path, destination,` or `source` returns only the cases where these arguments are considered adjunct, as opposed to core, arguments. This somewhat naturally filters out lexical verbs with motion semantics (where these arguments are numbered, core arguments) and leaves more “exceptional” usages of verbs in CAUSED-MOTION or INTRANSITIVE-MOTION constructions. Again, this returns a variety of false positives, which were manually filtered to arrive at the desired number of instances for each construction.

The collection of instances of the DITRANSITIVE and RESULTATIVE were by far the most challenging. There is no single semantic arugment that allows for locating these in the AMR corpus. As a result, we conducted detailed syntactic searches of a parsed version of COCA for the syntactic shape of the constructions described above; namely: NP V NP NP and NP V NP OBL. This results in a staggering number of false positives fitting the same syntactic shape but lacking the appropriate semantic roles associated with those syntactic slots. The false positives were manually filtered and we then leveraged the Penn Treebank syntactic patterns of the few true positives that we found to feed back into our search. In rounds of refining the search, we discovered that the syntactic treatment of resultatives in particular is extremely inconsistent, with the postverbal slots sometimes treated as a clause, a small clause, or individual phrasal arguments of the verb. This highlights the lack of compatability between some existing syntactically and semantically tagged resources and what is required to develop and test CxG-based research hypotheses.

5. Corpus Use Case: Probing LLMs

We leverage our CoGS corpus to evaluate our hypothesis that models are likely to be able to perform better on distinguishing instances of substantive constructions and perform more poorly on distinguishing instances of purely schematic constructions. We note that substantive constructions are associated with formal linguistic competencies, whereas schematic constructions encompass the functional linguistic capabilities necessary for effectively utilizing world knowledge, comprehension of social contexts, and reasoning, as they must be distinguished semantically as well as syntactically from other formally similar constructions.

5.1. Experimental Setup

In evaluating our hypothesis, we use GPT-3.5 and GPT-4. These models, among the largest currently available, are likely to serve as representative indicators of the capabilities of other LLMs. Our experiments are conducted via the OpenAI API endpoints, utilizing default hyperparameters, except for the temperature, which is set to 0 to ensure reproducibility.

We evaluate the extent to which models are capable of differentiating sentences as instances of various constructions using a prompt-based probing method. Concretely, after initial experiments to determine the most effective phrasing, we use the following prompt:

> From amongst the following sentences, extract the three sentences which are instances of the ConstructionName construction, as exemplified by the following sentence: Sentence. Output only the three sentences in three separate lines:

For more well-known and consistently described constructions, LLMs are likely to be able to extract the required information based solely on the name of the construction. Nonetheless, since constructions might be referred to using slightly different names, we also provide an example to ensure clarity regarding the specific construction in question. Crucially, we evaluate two variations of the above prompt: the first is where we use a randomly extracted CoGS sentence as the example Sentence, and the second is where we provide an exemplar CoGS Sentence that is carefully selected based on several factors. These criteria include the sentence being concise, meaning it's not substantially longer than the construction itself, and also representing a typical use case of that construction, as determined by a linguist. Table 4 in the Appendix provides the exemplar sentences used for this variant of the prompt. Given that the exemplar is likely to carry more information pertaining to the construction in question we expect this probing method to perform better than when we use a randomly selected sentence which is an instance of the construction.

The probes consist of six options from which models are required to extract three sentences. So as to ensure a fine-grained evaluation, we define a model's extraction of a sentence as 'successful' if the output sentence is present in the list of target sentences or if any of the target sentences are found within the generated output. Furthermore, we permit partial retrieval and assign one-third of a point for each correctly extracted sentence. This setup leads a random baseline accuracy of 50%.

LLMs possess the capability to perform tasks given only a handful of examples within their prompts. This capability, termed *in-context learning* (Brown et al., 2020), is believed by some to be responsible for LLMs' effectiveness (Lu et al., 2023). As such, we additionally conduct our experiments with and without the addition of a single task example within the prompt. We refer to these experimental settings as one-shot (incorporating one example) and zero-shot (lacking any examples), respectively.

5.2. Results and Discussion

Our results are detailed in Table 3. Overall, consistently across our probes and the different models we experiment with, we find that the ability of models to identify constructions reduces with the increase in schematicity. This observation firmly confirms our initial hypothesis that LLMs excel at formal linguistic tasks (distinguishing substantive constructions) while struggling with functional linguistic tasks (distinguishing schematic constructions). Therefore, these results confirm the significance of employing constructional information in probing and improving the performance of LLMs.

Surprisingly, however, we find that the use of an exemplar Sentence exhibiting a canonical construct in the prompt does not improve the performance of our probes. This is contrary to our expectation and is particularly interesting as it is consistent across both the models we evaluate. While further research is required in this regard, we believe that one of the reasons for this could be that a shorter exemplar is likely to prevent the model from generalizing effectively to identify other sentence which are instances of the construction. Furthermore, the exemplar may skew the models to attend to lexical semantics relating to a particular topic, as opposed to constructional semantics.

Finally, we also find that model performance *drops* between the zero-shot and the one-shot setting, but only when we evaluate sentences which are less schematic. This finding is contrary to what we might expect, as the inclusion of even a single example in the prompt substantially improves performance. We attribute this drop in performance to models' ability to recognize a construction based on its name, rather than a specific example, even if an exemplar. Interestingly, however, the performance in the one-shot setting increases as the degree of schematicity increases. This observation is particularly interesting, as it implies that models encounter multiple challenges when attempting to identify instances of schematic constructions.

Overall, our findings indicate that LLMs, despite their extensive parameter count and web-scale training data, struggle with to recognize more schematic constructions. Our dataset, which is the first to offer a spectrum of schematicity for comprehensive evaluation, combined with our experimental setup for fine-grained assessment, enables us to pinpoint more precise aspects of language interpretation that pose difficulties for LLMs. We note that the prompts employed in our study, although optimized using initial experiments, remain simplistic. Additionally, it's important to note that evaluating prompt effectiveness is not the same as directly probing the internal representations of models (Hu and Levy, 2023). Similarly, evaluating meta-linguistic information encoded in LLMs is not equivalent to evaluating how effectively a model can utilize this information in downstream tasks. Nonetheless, our baseline experiments presented in this work aim to highlight the significance of this dataset, with the exploration of these more detailed tasks left for future research.

Setting	CxN	Abstraction Level	Random Example Prompt GPT-3.5 (%)	Random Example Prompt GPT-4 (%)	Exemplar Prompt GPT-3.5 (%)	Exemplar Prompt GPT-4 (%)
zero-shot	Let Alone	Substantive	98.00	100	95.33	97.33
	Much Less		70.00	96.67	70.67	91.33
	Average of all Substantive		84.00	98.34	83.00	94.33
	Comparative Correlative	Partial	79.33	96.00	82.67	94.67
	Way Manner		66.67	90.00	62.00	82.67
	Conative		88.67	99.33	90.67	96.67
	Causative with		66.00	85.33	76.67	90.67
	Average of all Partial		75.17	92.67	78.00	91.17
	Resultative	Schematic	56.67	62.00	48.67	62.00
	Intransitive Motion		54.67	61.33	53.33	60.00
	Caused Motion		57.33	59.33	61.33	44.67
	Ditransitive		47.33	66.67	40.67	52.67
	Average of all Schematic		54.00	62.33	51.00	54.84
	Average of all zero-shot		68.47	81.67	68.20	77.27
one-shot	Let Alone	Substantive	88.67	100	96.00	97.33
	Much Less		72.67	95.33	68.00	92.00
	Average of all Substantive		84.00	97.67	83.00	94.67
	Comparative Correlative	Partial	73.33	92.67	86.00	94.00
	Way Manner		62.67	91.33	68.00	81.33
	Conative		84.67	98.67	88.00	96.67
	Causative with		65.33	100	70.67	92.67
	Average of all Partial		71.50	95.67	78.17	91.17
	Resultative	Schematic	52.00	60.67	56.00	57.33
	Intransitive Motion		54.67	64.00	55.33	63.33
	Caused Motion		60.00	68.67	58.00	67.33
	Ditransitive		63.33	82.00	67.33	77.33
	Average of all Schematic		57.50	68.84	59.17	66.33
	Average of all one-shot		67.73	85.33	71.33	81.93

Table 3: Results of the Example and Exemplar probes across the zero-shot (above dashed line) and one-shot (below dashed line) settings on both GPT-3.5 and GPT-4. The random baseline is 50%. Results for constructions are listed from substantive to schematic, revealing the trend of an overall decrease in performance related to an overall increase in schematicity. See text for details.

6. Conclusions and Future Work

In this work we introduce CoGS, the Construction Grammar Schematicity corpus, which is designed to probe the extent to which LLMs can identify constructions at different levels of schematicity. We consistently find that models, despite their size, perform significantly worse at identifying those constructions with increased schematicity. Furthermore, we observe that models exhibit an increased reliance on the examples provided in the one-shot setting to identify constructions at higher levels of schematicity. This suggests an implicit absence of such information within the models themselves.

Our probes are intentionally crafted to assess a model's capacity to extract information related to a construction from both its name and an example. In future research, we plan to develop distinct probes to separately evaluate these different aspects. Furthermore, we intend to investigate how these evaluations impact downstream performance by expanding our corpus to incorporate a downstream natural language inference task.

Our results, demonstrating clear support for our theoretically motivated hypothesis, underscore that constructional information can be use to enhance the performance of LLMs, by tapping into the theoretical underpinnings of cognitive linguistics and usage-based grammar. We have shown that these theories enable us to predict where LLMs will perform best, and where models may struggle to interpret natural language. We will carry these findings forward to continue to expand constructional resources in service of advances in NLP.

7. Ethics Statement

All data collected has been obtained with explicit permission from the original data owners or in strict adherence to the terms of use specified by the data source. The utilization of Large Language Models has been undertaken following meticulous dataset curation to minimize any potential environmental impact associated with their use.

8. Bibliographical References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer, and Nathan Schneider. 2018. Abstract meaning representation of constructions: The more we include, the better the representation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bastian Bunzeck and Sina Zarrieß. 2023. Entrenchment matters: Investigating positional and constructional sensitivity in small and large language models. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 25–37.

Joan Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, pages 711–733.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Noam Chomsky. 2014. *The minimalist program*.

Mark Davies. 2008. The corpus of contemporary american english (coca): 560 million words, 1990-present.

Ferdinand De Saussure et al. 1916. Nature of the linguistic sign. *Course in general linguistics*, 1:65–70.

Jonathan Dunn. 2017. Computational learning of construction grammars. *Language and cognition*, 9(2):254–292.

Jonathan Dunn. 2023. Exploring the constructicon: Linguistic analysis of a computational cxg. *arXiv preprint arXiv:2301.12642*.

Charles Fillmore, Paul Kay, and Mary O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–531.

Charles J Fillmore, Russell Lee-Goldman, and Russell Rhodes. 2012. The framenet constructicon. *Sign-based construction grammar*, (193):309–372.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Adele E. Goldberg. to appear. A chat about constructionist approaches and lmms. *Constructions and Frames*.

Thomas Hoffmann. 2022. *Construction Grammar*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Thomas Hoffmann and Graeme Trousdale. 2013. *The Oxford handbook of construction grammar*. Oxford University Press.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Ray Jackendoff. 2010. *Meaning and the lexicon: the parallel architecture 1975-2010*. OUP Oxford.

Ray S Jackendoff. 1992. *Semantic structures*, volume 18. MIT press.

Matt A Johnson and Adele E Goldberg. 2013. Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes*, 28(10):1439–1452.

Michael P Kaschak and Arthur M Glenberg. 2000. Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of memory and language*, 43(3):508–529.

Kevin Knight et al. 2020. Abstract Meaning Representation (AMR) Annotation Release 3.0 LDC2020T02. Web Download. Philadelphia: Linguistic Data Consortium.

Natalia Levshina and Kris Heylen. 2014. A radically data-driven construction grammar: Experiments with dutch causative constructions. *Extending the scope of Construction Grammar*, 54:17.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning?

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Florent Perek. 2018. Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, 14(1):65–97.

Giulia Rambelli, Emmanuele Chersoni, Philippe Blache, Chu-Ren Huang, and Alessandro Lenci. 2019. Distributional semantics meets construction grammar. towards a unified usage-based model of grammar and meaning. In *First International Workshop on Designing Meaning Representations (DMR 2019)*.

Anatol Stefanowitsch and Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. Cxgbert: Bert meets construction grammar. *arXiv preprint arXiv:2011.04134*.

Harish Tayyar Madabushi, Laurence Romain, Petar Milin, and Dagmar Divjak. 2024. Construction grammar and language models. In Mirjam Fried and Kiki Nikiforidou, editors, *The Cambridge Handbook of Construction Grammar*. Accepted for publication, to appear.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Michael Tomasello. 2006. Construction grammar for kids. *Constructions*.

Tim Veenboer and Jelke Bloem. 2023. Using collostructional analysis to evaluate bert's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12937–12951.

Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R Bowman. 2020. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *arXiv preprint arXiv:2010.05358*.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative. *arXiv preprint arXiv:2210.13181*.

Appendix A. CoGs Dataset

We provide a summary table of the ten constructions, along with their level of schematicity and any substantive words, as well as a form and meaning description of the construction. For the form description, we leverage the phonology and morphosyntax constructional templates of Hoffmann (2022).

Construction	**Schematic / Substantive**	**Meaning Description**	**Form Description**	**Example of Exemplar Prompt**
Let-Alone	Substantive - “let + alone” is frozen	Invokes a scale of an implicit property, where preceding slot A is on the lower end in comparison to following slot B	PHONOLOGY: /A1 lɛt$_{2}$ ɘ’loʊn$_{3}$ B$_{4/5}$ MORPHOSYNTAX: /XP$_{1}$ CONJ$_{2\text{-}3}$ XP$_{4/5}$	[[None of these arguments is notably strong,]$_{1}$ let$_{2}$ alone$_{3}$ [conclusive]$_{4}$]$_{5}$
Much-Less	Substantive - “much + less” is frozen	Invokes a scale of an implicit property, where preceding slot A is on the lower end in comparison to following slot B	PHONOLOGY: /A$_{1}$ mʌtʃ$_{2}$ lɛs$_{3}$ B$_{4/5}$ MORPHOSYNTAX: /XP$_{1}$ CONJ$_{2\text{-}3}$ XP$_{4/5}$	[[He has not yet been put on trial]$_{1}$ much$_{2}$ less$_{3}$ [found guilty.]$_{4}$]$_{5}$
Way-Manner	Partially Substantive - PRP$ + “way” + path PP	Indicates movement of an entity while performing the action of a manner verb	PHONOLOGY: /A$_{1}$ B$_{2}$ [C$_{3}$ weI$_{4}$]$_{5}$ D$_{6/7}$ Morphosyntax: [SBJ$_{1}$ [V$_{2}$ [PRON$_{3=1}$] way$_{4}$]$_{\text{OBJ5}}$ OBL:PP$_{6}$]$_{\text{VP}}$]$_{7}$	[[A middle-aged man]$_{1}$ eased$_{2}$ his$_{3}$ way$_{4}$ [into$_{6}$ the room.]]$_{7}$
Comparative-Correlative	Partially Substantive - “the” + comparative ADJ/ADV. . . + “the” + comparative ADJ/ADV	Specifies a cause-and-effect and/or temporal relation between the first comparative slot, and the second comparative slot	PHONOLOGY: /[ðə$_{1}$ A$_{2}$ B$_{3}$]$_{\text{C1}}$ [ðə$_{4}$ C$_{5}$ D$_{6}$]$_{\text{C2/7}}$ Morphosyntax: [[*the*$_{1}$ [Comparative Phrase]$_{2}$ REST-CLAUSE$_{3}$]$_{\text{C1}}$ [*the*$_{4}$ [Comparative Phrase]$_{5}$ REST-CLAUSE$_{6}$]$_{\text{C2}}$]$_{7}$	[The$_{1}$ more$_{2}$ [I studied]$_{3}$, the$_{4}$ less$_{5}$ [I understood]$_{6}$]$_{7}$
Conative	Mostly Schematic - “at” is fixed	Indicates an attempt to transfer force from an agent to a patient	PHONOLOGY: /A$_{1}$ B$_{2}$ [æt $_{\text{C}}$]$_{3/4}$ Morphosyntax: [SBJ$_{1}$ [V$_{2}$ OBL:*at*-PP$_{3}$]$_{\text{VP}}$]$_{4}$	[[She]$_{1}$ kicked$_{2}$ [at the ball] $_{3}$]$_{4}$
Causative-With	Mostly Schematic - “with” is fixed	Denotes ground as an affected patient that ends up in a particular state as the result of the application of a theme	PHONOLOGY: /A$_{1}$ B$_{2}$ C$_{3}$ [wIð D]$_{4/5}$ MORPHOSYNTAX: [SBJ$_{1}$ [V$_{2}$ OBJ$_{3}$ OBL:*with*-PP$_{4}$]$_{\text{VP}}$]$_{5}$	[[She]$_{1}$ loaded$_{2}$ [the truck]$_{3}$ [with$_{4}$ books]$_{4}$]$_{5}$

Caused-Motion	Fully Schematic	Agent of the action denoted by the verb causes theme to move along or towards a goal	PHONOLOGY: /A_1 B_2 C_3 $D_{4/5}$ MORPHOSYNTAX: [SBJ_1 [V_2 OBJ_3 OBL_4]$_{VP}$]$_5$	[[They]$_1$ laughed$_2$ [the actor]$_3$ [off the stage]$_4$]$_5$
Intransitive-Motion	Fully Schematic	A theme carries out an event that causes or accompanies movement	PHONOLOGY: /A_1 B_2 $C_{3/4}$ MORPHOSYNTAX: [SBJ_1 [V_2 OBL_3]$_{VP}$]$_4$	[[The fly]$_1$ buzzed$_2$ [into the room]$_3$]$_4$
Ditransitive	Fully Schematic	Agent of the action denoted by the verb is construed as (intending to) cause a recipient to receive a theme.	PHONOLOGY: /A_1 B_2 C_3 $D_{4/5}$ MORPHOSYNTAX: [SBJ_1 [V_2 OBJ_3 OBJ_4]$_{VP}$]$_5$	[[She]$_1$ baked$_2$ [her sister]$_3$ [a cake]$_4$]$_5$
Resultative	Fully Schematic	Agent of the action denoted by the verb causes a patient to change / become a resulting state	PHONOLOGY: /A_1 B_2 C_3 $D_{4/5}$ MORPHOSYNTAX: [SBJ_1 [V_2 OBJ_3 OBL_4]$_{VP}$]$_5$	[[Firefighters]$_1$ cut$_2$ [the man]$_3$ free$_{45}$

Table 4: Ten constructions of focus in this research, which we use to collect a corpus and evaluate LLMs for access to constructional information. The Form pole description leverages constructional templates to distinguish phonologically fixed, substantive slots (denoted in IPA), and the morphosyntactic character of schematic slots (denoted with variables in the phonological description that are co-indexed with grammatical constituent types in the morphosyntactic layer). The examples given here are the exemplar sentences provided in the *Exemplar* probe.