CxGsNLP 2025

The Second International Workshop on Construction Grammars and NLP

Proceedings of the Workshop

September 24, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-318-0

Introduction

Constructionist approaches to language posit that all linguistic knowledge needed for language comprehension and production can be captured as a network of form-meaning mappings, called constructions. Construction Grammars (CxGs) do not distinguish between words and grammar rules, but allow for mappings between forms and meanings of arbitrary complexity and degree of abstraction. CxGs are thereby able to uniformly capture the compositional and non-compositional aspects of language use, making the theory particularly attractive to researchers in the field of Natural Language Processing (NLP). CxG theories, for example, can serve as a valuable 'lens' to assess and investigate the abilities of today's large language models, which lack explicit, theoretically grounded linguistic insights. At the same time, techniques from the field of NLP are often employed for the further development and scaling of CxG theories and applications.

The inaugural Construction Grammars (CxGs) and Natural Language Processing (NLP) (CxGs+NLP) workshop[1] successfully initiated dialogue between the two complementary perspectives of CxG and NLP, highlighting the untapped potential for collaboration and knowledge exchange. The first workshop took place shortly after the release of ChatGPT; now, two years later, the field has advanced considerably with the rise of generative AI and new LLMs. These developments make it all the more compelling to bring together researchers and practitioners to discuss the evolving landscape of CxG and NLP. In addition, in the time since the first workshop, there has been significant growth in the community's interest at this intersection. Building on this momentum, the second CxGs+NLP workshop brings together researchers across theory and practice once again to explore how CxG approaches can both inform and benefit from state of the art NLP methods, with an emphasis on LLMs.

These proceedings include papers presented at the 2nd International Construction Grammars and NLP workshop on 24 September 2025, held in conjunction with 16th International Conference on Computational Semantics (IWCS) in Dusseldorf, Germany. CxGs+NLP 2025 received 35 submissions, out of which 17 archival presentations were presented as in-person talks, 5 papers were presented virtually as lightning talks with posters, and 9 non-archival papers were presented during the in-person poster session. The papers address topics including computational frameworks and tools for CxG, LLMs, constructional knowledge and evaluation, and empirical studies and theoretical insights.

In addition to the oral paper presentations and poster session, CxGs+NLP 2025 featured three outstanding invited talks by Professor Adele Goldberg (Psychology, Princeton University), Professor Laura A. Michaelis (Linguistics, University of Colorado Boulder), and Professor Thomas Hoffmann (English Language and Linguistics, Catholic University of Eichstätt-Ingolstadt).

Our program also included a second, community-building day of events on 25 Sept 2025. This event featured panels and break-out sessions to spur discussion and development of persistent community resources and points for communication and data-sharing. We encourage readers to join our community by joining the online CxGs+NLP Group, which we continue to maintain with outcomes of our workshops and upcoming events.

Message from the Workshop Chairs

We thank our organizing committee for its continuing organization of the CxGs+NLP workshops, and the IWCS 2025 workshop chairs for their support. We are grateful to all of the authors for submitting their papers to the workshop and our program committee members for their dedication and their thoughtful reviews. We thank our invited speakers for making the workshop a uniquely valuable discussion of CxGs+NLP research.

Claire Bonial, Harish Tayyar Madabushi

[1] https://sites.google.com/view/cxgsnlpworkshop

Organizing Committee

General Chairs

Claire Bonial, Georgetown University and Army Research Lab
Harish Tayyar Madabushi, University of Bath

Organizing Committee

Melissa Torgbi, University of Bath
Leonie Weissweiler, Uppsala University
Austin Blodgett, Army Research Lab
Katrien Beuls, Université de Namur
Paul Van Eecke, Vrije Universiteit Brussel

Program Committee

Program Chairs

Katrien Beuls, Université de Namur
Austin Blodgett, Army Research Lab
Claire Bonial, Georgetown University and Army Research Lab
Harish Tayyar Madabushi, University of Bath
Melissa Torgbi, University of Bath
Paul Van Eecke, Vrije Universiteit Brussel
Leonie Weissweiler, Uppsala University

Reviewers

Katrien Beuls, Jérôme Botoko Ekila, Gosse Bouma, Bastian Bunzeck, Miriam Butt

Ming Cai

Liesbet De Vos, Stefania Degaetano-Ortlieb, Soumik Dey, Lucia Donatelli, Jonathan Dunn

Kilian Evang

Loïc Grobol

Stefan Hartmann, Dag Trygve Truslew Haug

Julia Kuznetsova

Alessandro Lenci, Olga Lyashevskaya

Alexander Mehler

Joakim Nivre

Timothy John Osborne, Rainer Osswald, Robert Östling

Laura Patrizzi

Mathilde Regnault, Laurence Romain

Wesley Scivetti

Ashwini Vaidya, Paul Van Eecke, Lara Verheyen, Remi van Trijp

Yuri V. Yerastov

Eva Zehentner

Table of Contents

A Computational Construction Grammar Framework for Modelling Signed Languages

Liesbet De Vos
Faculté d'informatique
Université de Namur
Belgium
liesbet.devos@unamur.be

Paul Van Eecke
Artificial Intelligence Laboratory
Vrije Universiteit Brussel
Belgium
paul@ai.vub.ac.be

Katrien Beuls
Faculté d'informatique
Université de Namur
Belgium
katrien.beuls@unamur.be

Abstract

Constructional approaches to signed languages are becoming increasingly popular within sign language linguistics. Current approaches, however, focus primarily on theoretical description, while formalization and computational implementation remain largely unexplored. This paper provides an initial step towards addressing this gap by studying and operationalizing the core mechanisms required for representing and processing manual signed forms using computational construction grammar. These include a phonetic representation of individual manual signs and a formal representation of the complex temporal synchronization patterns between them. The implemented mechanisms are integrated into Fluid Construction Grammar and are available as a module within the Babel software library. Through an interactive web demonstration, we illustrate how this module lays the groundwork for future computational exploration of constructions that bidirectionally map between signed forms and their meanings.

1 Introduction

Constructional approaches to signed languages are becoming increasingly popular within sign language linguistics (see Wilcox and Martínez (2025) for an overview). One possible reason for this popularity is construction grammar's potential to address key challenges in the field. The view on lexicon and grammar as a continuum, for instance, can help resolve longstanding problematic distinctions between grammatical and lexical signs on the one hand (Lepic and Occhino, 2018; Lepic, 2019), and gestural and linguistic signs on the other hand (Lepic and Occhino, 2018; Occhino and Wilcox, 2017; Wilcox and Xavier, 2013).

Despite its popularity, construction grammar has primarily been used for the theoretical description of sign language constructions (e.g., Lepic and Occhino, 2018; Schembri et al., 2018; Lepic, 2019; Wilcox and Occhino, 2016; Hou, 2022a,b; Wilcox and Martínez, 2020; Martínez et al., 2019; Wilcox et al., 2022; Johnston and Ferrara, 2012) while formalization and computational implementation of these constructions remain relatively unexplored. One exception is the work by van Trijp (2015), who provides a proof-of-concept implementation of two French Sign Language (LSF) constructions. Other computational models rely on formalisms such as Head-Driven Phrase Structure Grammar (HPSG) (Elliott et al., 2008), Role and Reference Grammar (RRG) (Murtagh, 2011b) or sign language specific production rules (Filhol et al., 2017). Except for van Trijp (2015)'s bidirectional approach, most existing computational work focuses solely on sign language production through avatar systems.

This paper presents an initial step towards addressing these gaps by studying and operationalizing the core mechanisms needed for representing and processing signed languages bidirectionally using a computational construction grammar framework. The development of such a framework is beneficial as it allows linguistic hypotheses to be formalized and verified on large linguistic corpora (van Trijp et al., 2022).

The implemented mechanisms include a phonetic, language-agnostic representation of individual manual signs and a formal description of the complex temporal synchronization patterns between them. These mechanisms are integrated into Fluid Construction Grammar (FCG) (Steels, 2011; Beuls and Van Eecke, 2023; van Trijp et al., 2022), a computational construction grammar framework that operationalizes the basic tenets of construction grammar (Steels, 2011; van Trijp et al., 2022; Beuls and Van Eecke, 2023, 2025). The developed framework is available as a module within the Babel software library [1], a toolkit containing all

[1] Download information is available at: https://emergent-languages.org

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 1–12

necessary modules for constructional language processing using FCG (Nevens et al., 2019; Loetzsch et al., 2008). While this module is an initial step towards bidirectional processing of sign language, more work is needed to scale our approach towards different research contexts and large-scale corpora.

To demonstrate the use of the framework, we provide an interactive web demonstration alongside this paper. It illustrates how a French Belgian Sign Language (LSFB) utterance can be comprehended and produced, showcasing the potential of our framework for future computational exploration of sign language constructions.

The remainder of this paper is structured as follows. First, we review existing representation systems (Section 2) and computational models (Section 3) for signed languages. Then, we introduce the implemented mechanisms for representing and processing manual signed forms, which are available as a module within the Babel framework (Section 4). Afterward, we illustrate the use of this module through an interactive web demonstration (Section 5). Finally, we conclude the paper (Section 6).

2 Representing signed forms

Accurately representing signed forms is one of the main challenges in developing computational models of sign language. Sign language expressions include movements produced by the entire upper body, including manual (hands) and non-manual articulators (face, head, eyes, shoulders, etc.). These movements can overlap in time, resulting in simultaneous/multilinear structures which cannot easily be captured using linear representation systems (Huenerfauth, 2006; Filhol and Braffort, 2012; Filhol, 2012). In addition, signers make extensive use of the three-dimensional space in front of the upper body to introduce and manage referents (Wilcox and Martínez, 2020). As a result, representations should include a fine-grained model of this three-dimensional space. Despite the challenges involved, several types of representation exist, including video, glosses, formal notation systems and avatar-specific representations.

The most frequently used format for representing signed expressions is video. It accurately captures the simultaneous nature of signing and is relatively easy to collect. However, it generally needs to be complemented with additional information for the purpose of linguistic description/modelling (Crasborn, 2015). Glosses are lexical labels which describe the prototypical meaning of a sign using words from the ambient spoken language. They do not provide any information about the internal structure of a sign and focus primarily on manual activity. In contrast, formal notation systems such as SignWriting (Sutton, 1995) or HamNoSys [2] (Prillwitz et al., 1987; Hanke, 2004) describe the sublexical structure of signs using a set of iconic symbols. This sublexical structure is typically described using a set of manual (hand shape, orientation, location and movement), and non-manual parameters (e.g. eye-gaze, brow, shoulder or head movements, etc.). Finally, avatar-specific representations capture the sublexical structure of signs from an animation perspective rather than a linguistic one, resulting in detailed descriptions of joint positions and rotations (Naert et al., 2020).

With the exception of video-based and some avatar-specific representations, most systems focus on describing isolated signs. Although these descriptions often can be concatenated, they fail to accurately represent the multilinearity of continuous signed expressions. For instance, while HamNoSys allows concatenation of individual sign descriptions to represent utterances, it lacks the capacity to represent instances where articulators act independently from each other, each producing forms with different start and end times (Filhol, 2012; Filhol and Braffort, 2012).

In linguistic research, this issue is addressed using multilayered annotation tools such as ELAN (Crasborn and Sloetjes, 2008; Dreuw and Ney, 2008) and ILEX (Hanke and Storz, 2008). They enable multiple annotation tiers to be aligned with a single time track, often derived from a video recording. Using this methodology, one layer can be created for each articulator and aligned to the time track of the video. Temporal relationships are conveyed implicitly through the start and end times of the recorded segments.

Another approach has been explored within the field of avatar synthesis, where systems such as the partition/constitution (Huenerfauth, 2004) and AZalee model (Filhol and Braffort, 2012; Filhol, 2012) explicitly describe temporal relationships between articulators. The partition/constitution model represents the structure of signed utterances through hierarchical syntax trees, where nodes branch into child nodes using constitution or par-

[2] Hamburg Notation System

tition (Huenerfauth, 2004). Constitution denotes traditional sequential branching, while partition accounts for simultaneous production. This hierarchical framework enables the modelling of complex synchronisation patterns across multiple articulators. However, synchronisation is constrained by the syntactic structure of the utterance. In contrast, the AZalee model supports a more flexible representation, allowing for the free arrangement of articulator movements by defining temporal relationships directly between their start and end boundaries (Filhol and Braffort, 2012; Filhol, 2012).

3 Computational Models of Sign Language

Computational models of sign language are relatively scarce, especially those that handle both language production and comprehension. While some work has focussed on producing sign language utterances using a grammatical model, little to no attention has been paid towards comprehending the semantic structure of sign language expressions. For production, several grammatical frameworks have been used to support avatar synthesis, including Head-Driven Phrase Structure Grammar (HPSG), Role and Reference Grammar (RRG) and AZee Production rules. For bidirectional processing, van Trijp (2015) explored the use of FCG.

The ViSiCAST and eSIGN projects use HPSG (Pollard and Sag, 1994) within a translation system from English to British Sign Language (BSL) (Elliott et al., 2008). The translation process involves multiple steps, including the use of BSL-specific HPSG rules that transform a semantic Discourse Representation Structure (DRS) into HamNoSys format (Elliott et al., 2008; Marshall and Sáfár, 2004, 2002; Sáfár and Glauert, 2010). The grammar includes 50 lexical and 9 grammatical rules and handles aspects like space, plurality, sentence types, and pronominal reference (Elliott et al., 2008; Marshall and Sáfár, 2004, 2002; Sáfár and Glauert, 2010). The HamNoSys output can be rendered into an XML-format which drives avatar animation (Kennaway, 2004).

Murtagh (2011b,a) first explored the use of RRG (Van Valin Jr. and Foley, 1980; Van Valin Jr., 1992) as a grammatical model to drive animation of Irish Sign Language (ISL). RRG is a functional theory which focusses primarily on the relationship between semantic, pragmatic, and syntactic structure (Van Valin Jr. and Foley, 1980; Van Valin Jr., 1992). The RRG grammar for ISL maps a semantic representation to a syntactical structure which represents manual, non-manual and temporal information (Murtagh et al., 2022). It can be transformed into a more detailed format which drives avatar animation (Murtagh et al., 2022). Amongst other aspects, the RRG approach focusses on modelling different verb types of ISL (Murtagh, 2020), including so-called directional/agreement verbs, which use the signing space to refer to arguments. While bidirectional, RRG has mainly been applied to sign language production (Murtagh et al., 2022).

AZee avoids spoken language categories such as noun, verb or adjective and describes sign language grammar using a set of production rules (Filhol et al., 2017). Each rule maps a semantic function to a score of time-aligned articulator movements. The output of one rule can serve as an argument to another, creating complex structures which drive avatar animation (Filhol et al., 2017). AZee rules have mostly been developped for French Sign Language (LSF), focussing on a wide range of linguistic phenomena, including interrogation (Martinod and Filhol, 2024), non-manual gestures (Challant and Filhol, 2024; Filhol et al., 2014), usage of space (Filhol and McDonald, 2022), and plurality (Martinod et al., 2022). The rules developped for LSF achieve 94% coverage on a moderately sized corpus (Challant and Filhol, 2022) and ongoing work seeks to expand this coverage (Challant and Filhol, 2024; Martinod et al., 2022).

Finally, van Trijp (2015) explores the use of computational construction grammar for sign language processing. He provides a proof-of-concept implementation of two LSF constructions using the FCG framework: a construction that handles the modification of sign parameters (i.e., hand shape, orientation, movement, location) to alter the meaning of a sign, and a construction which deals with the inherent multilinearity of continuous signed expressions (van Trijp, 2015). In contrast to other approaches discussed in this section, the FCG approach proposed by van Trijp (2015) is bidirectional, allowing language production and comprehension. The proof-of-concept implementations show the potential of computational construction grammar (specifically FCG) for modeling signed languages.

4 A Computational Construction Grammar Framework for Modelling Signed Languages

The main objective of the proposed framework is to support the computational exploration of sign language constructions in comprehension and production. To achieve this, we identify three core properties:

1. **Phonetic representation:** The framework should accurately represent the realisation of manual sign forms, including use of three dimensional space and the shape, orientation, location and movements of the hands.

2. **Multilinear representation:** The framework should provide explicit temporal relations between manual articulator movements, capturing the inherent multilinear nature of signed languages.

3. **Bidirectional:** The framework should allow bidirectional processing between signed forms and their meanings.

The remainder of this section describes how we integrate each of these properties into the proposed framework for sign language processing and release it as a module within the Babel framework.

4.1 HamNoSys: Phonetically Representing Signed Forms

Before 1960, there was an overall consensus that sign language forms lacked internal structure, making them unfit for linguistic study. This consensus changed after William Stokoe published his Pioneering work on American Sign Language (ASL) phonology in 1960 (Stokoe, 1960). Stokoe argued that, similar to spoken forms, signed forms have internal structure, determined by three main parameters: hand configuration (including the shape and orientation of the hand), location and movement (Stokoe, 1960). To support his claims, Stokoe identified minimal pairs in ASL, where only one of these three parameters distinguishes the two forms, illustrating their contrasting ability within the language. While Stokoe's work was phonological, his theory fuelled many phonetic theories and writing systems for signed forms. These often describe the structure of the sign using Stokoe's basic parameters (handshape, orientation, location and movement). Modern phonetic theories and writing systems also include non-manual features (facial expressions or shoulder and head movements).

HamNoSys (Hamburg Notation System) is one of the writing systems building on Stokoe's theory for sign structure. It relies on a set of iconic glyphs to represent hand shape, orientation, location, and movement, along with non-manual components. It is a phonetic alphabet that was created as a sign language counterpart to the International Phonetic Alphabet[3](IPA). Like IPA, it contains symbols that can be used to describe the phonetic realisation of signs in any sign language. HamNoSys is well integrated with modern computer software, having a Unicode font and XML-based representation known as Signing Gesture Markup Language (SiGML) (Elliott et al., 2004). Through SiGML, HamNoSys strings can be converted into avatar animations using the JASigning system[4].

The basic structure of a HamNoSys representation consists of two core components: an initial posture and a set of actions (see Figure 1). The initial posture describes the shape, orientation, and location of the hands at the onset of the sign. Actions describe movements through space that might change part of the initial structure. For two-handed signs, the initial posture is preceded by a symmetry operator. This operator specifies how the non-dominant hand copies information from the dominant hand, avoiding repetition. For a more detailed description, we refer to the HamNoSys manual (Smith, 2013). Figure 1 illustrates an example of an HamNoSys representation for the sign RIVIERE (RIVER) in French Belgian Sign Language (LSFB).

We choose HamNoSys as a phonetic representation as it is language agnostic, well-documented and -integrated into modern computer software (Hanke, 2004). It contains a fine-grained model of the signing space, which is crucial for grammatical modelling. Reading and writing HamNoSys requires some initial training, but the existance of an extensive manual (Smith, 2013), unicode font, input palettes (see Hanke, 2021), and avatar software make the system easy to learn and enjoyable to use.

[3]For more information, see: https://www.internationalphoneticassociation.org

[4]more information about the JASigning system: https://vh.cmp.uea.ac.uk/index.php/JASigning

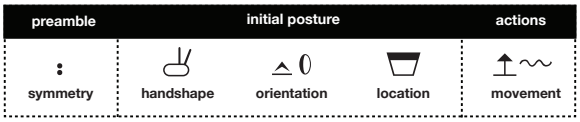


Figure 1: Example of a HamNoSys representation for the LSFB sign RIVIERE (RIVER). The sign is produced using two hands that behave symmetrically. The initial posture and actions are only described for the dominant hand, as those for the non-dominant hand can directly be inferred from them. The dominant hand's shape is a fist with the index and middle fingers extended. Fingers are oriented outwards, and the palm is directed to the left of the signer (right for the non-dominant hand). Hands move outward using a wavy motion.

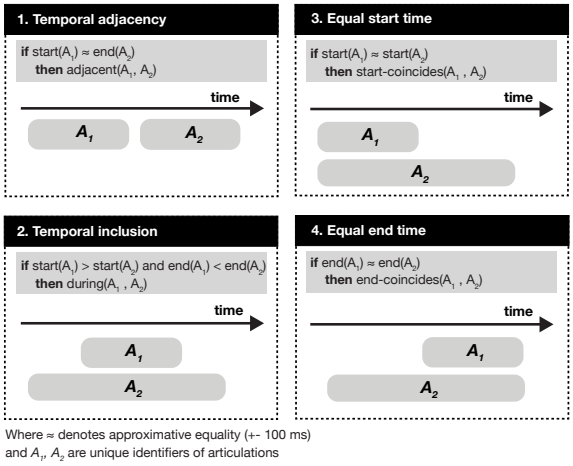


Figure 2: The collection of temporal relationships used within the multilinear representation of our framework. The adjacency relationship captures the sequential ordering of two articulations, while the remaining relationships describe multilinear structures involving two articulations. A single articulation can be involved in multiple relationships.

4.2 Using Temporal Relationships to Convey Multilinear Structure

To describe the temporal relationships between manual signs, we first use the ELAN annotation software to align the two manual articulators to a single time-track, derived from the video-recording of the expression. Annotation files contain two layers for each of the manual articulators (four layers total). The first layer for each hand divides the time-track into individual signs, identifying each sign using its lexical ID-gloss label. The second layer provides a HamNoSys representation for each identified sign.

Afterward, the annotation file is used to extract the ID-gloss, HamNoSys representation, and temporal boundaries for each identified segment. In our multilinear representation, every articulation is modeled as a predicate that specifies its type (i.e. *two-hand-articulation*, *right-hand-articulation* or *left-hand-articulation*), and takes two arguments: a unique identifier (derived from the sign's ID-gloss) and a HamNoSys string. For every pair of articulations, we evaluate whether any of the temporal relationships illustrated in Figure 2 apply. The resulting output is a set of predicates that encode both the phonetic structure of individual manual signs and the temporal relations between them.

Figure 3 shows a multilinear representation of an LSFB question from the GeoQuery-LSFB corpus[5]. The LSFB expression is a translation of the English question *"What are the high points of states surrounding Mississippi?"*. To illustrate the use of temporal relationships, we focus on the highlighted part of the expression. The left hand produces a sign glossed as DS[BENT5]:ETAT, which depicts a state. Meanwhile, the right hand performs three sequential signs: DS:[BENT5]:ETAT+, depicting multiple states, PT:DET/LOC[1]+, a pointing sign referring to the locations of the previously introduced states, and HAUT, which refers to a high point. The multilinear representation specifies each articulation's type, unique identifier and HamNoSys representation, along with five temporal relationships: two adjacency relations between the right handed signs, an equal start relationship between DS[BENT5]:ETAT+ and DS[BENT5]:ETAT, a during relationship between PT:DET/LOC[1]+ and DS[BENT5]:ETAT, and an equal end relationship between HAUT and DS[BENT5]:ETAT.

Our multilinear representation was primarily inspired by the AZalee approach, where articulators from any type can be aligned freely through their temporal boundaries and a set of temporal relationships (Filhol, 2012; Filhol and Braffort, 2012). While our representation currently only includes manual articulators, we acknowledge the importance of non-manual components for grammar modelling. Therefore, we designed the representation to be extensible, aiming to add non-manual articulation types in the future.

[5]A resource of LSFB questions on U.S. geography, annotated with procedural semantic representations. Available at: https://gitlab.unamur.be/beehaif/GeoQuery-LSFB

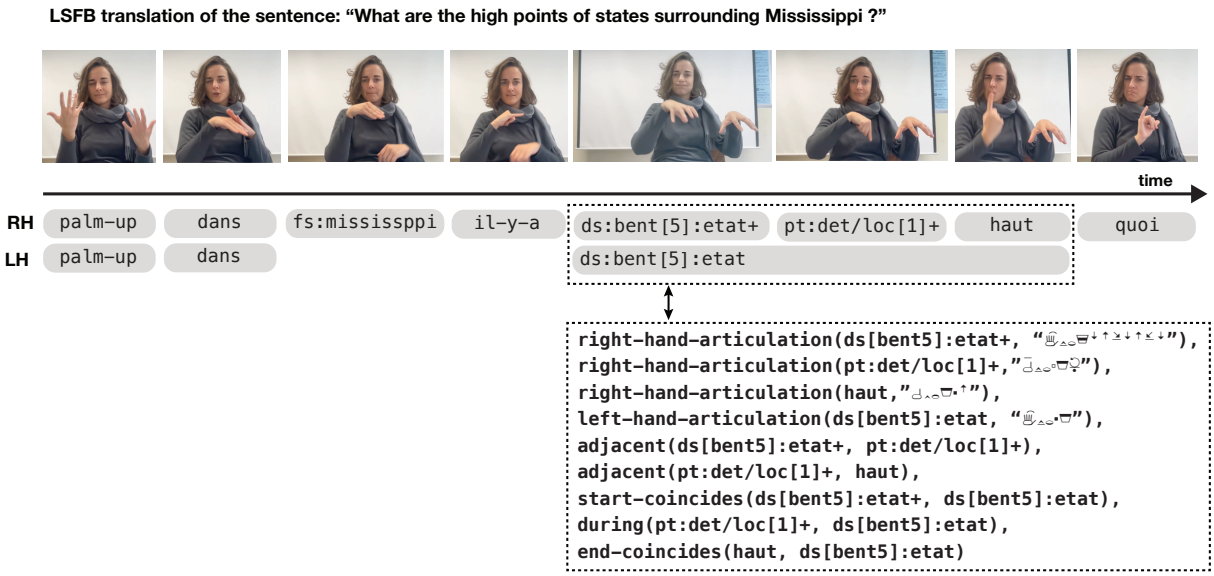


Figure 3: Example of a multilinear representation for a form in LSFB. The highlighted part of the expression contains four articulations in total, three right handed and one left handed one. Right handed articulations all occur in sequence (conveyed by two adjacency relationships), while the left handed articulation is held through time and simultaneously occurs with the sequence of right handed signs. This simultaneous occurrence is conveyed through the start-coincides, during, and end-coincides relationships in the representation.

4.3 Fluid Construction Grammar: a Bidirectional Computational Construction Grammar Framework

The proposed mechanisms for processing signed languages are integrated into the well-established FCG framework (Steels, 2011; van Trijp et al., 2022; Beuls and Van Eecke, 2023, 2025). It models language processing as a state-space search process (see Figure 4), with the initial state containing the input (form in comprehension, meaning in production) and the goal state containing an expanded structure capturing both the form and meaning of the utterance (Van Eecke et al., 2022; Van Eecke and Beuls, 2017). Between start and end states, constructions or form-meaning mappings act as operators that expand this initial state (Van Eecke et al., 2022). Representations of intermediate states within this process are commonly referred to as transient structures. The entire state-space search process is referred to as the construction application process. For more information about constructional language processing using FCG, please consult Van Eecke and Beuls (2017) and Van Eecke et al. (2022).

Within the FCG sign language module, linguistic forms are represented using our multilinear representation, while any formal semantic framework can be used to represent their meaning. Examples are procedural semantics (Woods, 1968; Woods et al., 1972; Winograd, 1972), discourse representation structure (Kamp and Reyle, 2013), and abstract/uniform meaning representation (Banarescu et al., 2013; Van Gysel et al., 2021). Like the general FCG architecture, grammars developed using the sign language package allow bidirectional processing, meaning they can be used to map form to meaning (comprehension) and meaning to form (production).

4.4 The FCG module for processing signed languages

We bundle the implemented mechanisms as an FCG module integrated into the Babel software library, allowing researchers to use and test the developed framework on their own sign language data.

The module contains several components, including methods for reading in ELAN annotation files and transforming them into the multilinear predicate notation format. A second component integrates these multilinear forms with the FCG construction application process, adding them to the initial transient structure in comprehension and extracting them from the final transient structure in production. Finally, a visualisation component integrates avatar animations and graphical repre-

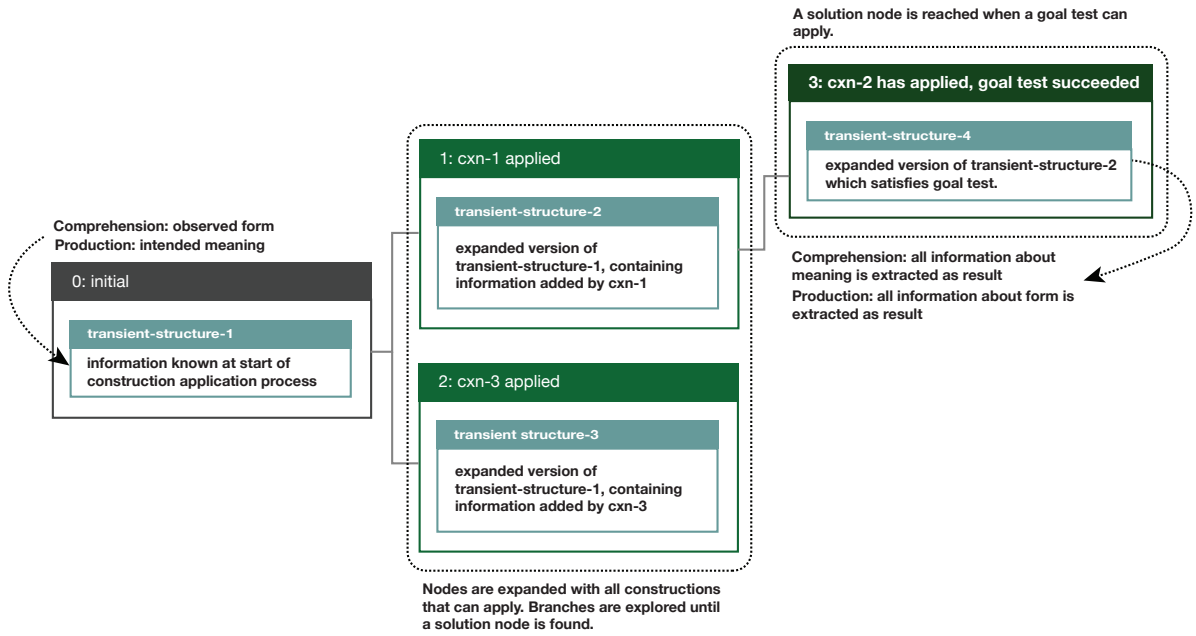


Figure 4: Illustration of the construction application process in FCG. The initial state (also referred to as initial transient structure), contains all information known at the start of processing (form in comprehension, meaning in production). Nodes are expanded by constructions and branches are explored until a solution node is reached.

sentations of multilinear structures, allowing users to inspect signed forms visually.

To use the FCG sign language processing module, a recent installation of the Babel toolkit is required (see the Babel installation page). After installing Babel and following the steps within the README of its sign language processing module, users can start using the framework.

5 Interactive Web Demonstration

To demonstrate the potential of the developed framework for the computational exploration of sign language constructions, we provide an interactive web demonstration[6] alongside this paper. It illustrates the functionality of our framework through the comprehension and production of the LSFB expression from Figure 3. The visualisations shown within the web demonstration are integrated into the FCG module for sign language processing as live visualisations, allowing users to inspect the construction application process for signed utterances in real time.

Figure 5 provides a schematic overview of the comprehension process. The initial transient structure contains the predicate notation of the observed form. Nodes in the search tree are expanded until a goal state with a complete meaning analysis is found. The transient structure of this final node contains a collection of units that were created by the applied constructions and combines information about the utterance's form and meaning. To complete the comprehension process, the meaning predicates are extracted from the final transient structure, resulting in a coherent meaning representation for the observed form.

The figure shows two constructions in detail: a concrete construction that captures the form and meaning of the LSFB sign HAUT (HAUT-CXN) and a more abstract construction which captures the typical theme-question format of questions in LSFB (THEME-QUESTION-CXN). The HAUT-CXN maps between the linguistic form (right-handed articulation with the index finger pointing upwards and performing a slight upwards movement) and the meaning of the sign HAUT (procedural meaning referring to a high point). The construction captures additional linguistic information about the sign, such as its semantic class, number, and location. The *args* feature later connects the sign's semantic arguments to those of other constructions. The THEME-QUESTION-CXN captures the information structure of LSFB questions, where the theme precedes the querying component. It enforces this precedence through an adjacency relation between the final sign of the theme and the first sign of the

[6]Available here: https://liesbet-devos.github.io/SL-processing-demo/

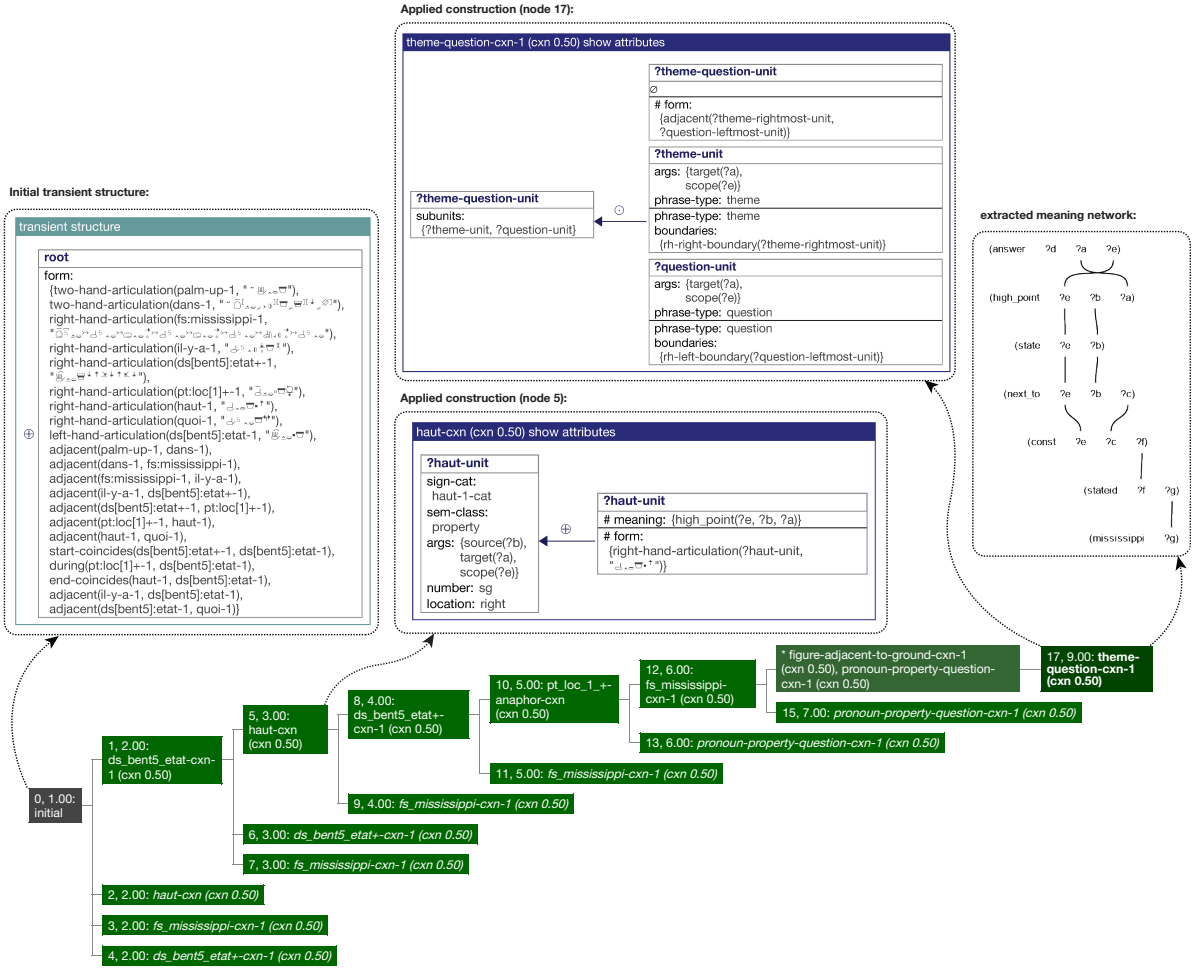


Figure 5: Schematic representation of the comprehension process for the signed expression of Figure 3. The initial transient structure contains the observed form in predicate format. A goal state is reached in node 17. To complete the comprehension process, all meaning predicates are extracted from the final transient structure, resulting in the meaning representation which is shown. During the application process, multiple constructions apply, amongst which the HAUT and THEME-QUESTION construction.

question. When this adjacency condition is met, the construction links the semantic arguments of both parts, resulting in a coherent meaning network for the expression.

This demonstration showcases how the FCG module for sign language processing facilitates comprehension and production of an LSFB expression. Its broader aim is to showcase the potential of FCG and the implemented sign language processing mechanisms for future computational exploration of sign language constructions.

6 Conclusion

The main goal of the current paper was to study and operationalize the core mechanisms required for representing and processing signed languages using computational construction grammar. We identified three core properties for the framework. It should (1) include a phonetic representation for manual signs, (2) make temporal relationships between these signs explicit, and (3) allow bidirectional processing. For phonetic representation, we rely on the well-established HamNoSys system, which describes the hand shape, orientation, location, movement and non-manual features of isolated signs. It is language-agnostic and represents signs from any sign language. While it provides the possibility to describe non-manual components as well, we do not yet include these in our approach, leaving this to future work. To describe temporal relationships between these signs, we propose a multilinear representation which extracts temporal information from ELAN annotation files. To allow bidirectional processing, we integrate the

developed mechanisms into the FCG framework. The implemented mechanisms are available as a module within the Babel software library, which is openly available.

Through an interactive web-demonstration, we illustrate how the proposed mechanisms effectively represent and process the multilinear forms of a signed language. More broadly, this demonstration showcases the potential of the proposed framework for future computational exploration of sign language constructions. While it is an initial step towards a functional framework for bidirectional sign language processing, challenges remain before our approach can be scaled to large corpora and various research contexts. An example is the inclusion of non-manual components and their temporal relations.

Limitations

A considerable limitation of the presented module is its focus on manually signed forms. Non-manual features are frequent within sign language productions and play crucial roles in many sign language constructions. However, formalising these non-manual forms remains challenging, with most formal notation systems focusing primarily on manual forms. Systems like HamNoSys often include some non-manual features, but they are not as extensive and well-developed as the manual ones. Another limitation of the framework is the collection of temporal relationships, which currently does not capture differences in on- or offset time between two articulations. Such differences might be needed to capture constructions that contain non-manual features, which often have a different on- or offset time. Finally, we have only tested the module on LSFB examples. While we expect the system to apply to other sign languages, it remains to be verified empirically.

Acknowledgements

We would like to acknowledge Lara Verheyen and the anonymous reviewers for their valuable comments on earlier versions of this manuscript. The research presented in this paper was made possible by the ARIAC research project (grant number 2010235), funded by DigitalWallonia4.ai.

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.

Katrien Beuls and Paul Van Eecke. 2023. Fluid Construction Grammar: State of the art and future outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 41–50. Association for Computational Linguistics.

Katrien Beuls and Paul Van Eecke. 2025. Construction grammar and artificial intelligence. In Mirjam Fried and Kiki Nikiforidou, editors, *The Cambridge Handbook of Construction Grammar*, pages 543–571. Cambridge University Press, Cambridge, United Kingdom.

Camille Challant and Michael Filhol. 2022. A first corpus of azee discourse expressions. In *Proceedings of the 13th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Camille Challant and Michael Filhol. 2024. Extending azee with non-manual gesture rules for french sign language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7007–7016. ELRA and ICCL.

Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 39–43. European Language Resources Association (ELRA).

Onno A. Crasborn. 2015. *Transcription and Notation Methods*, chapter 5. John Wiley & Sons, Ltd.

Philippe Dreuw and Hermann Ney. 2008. Towards automatic sign language annotation for the elan tool. In *sign-lang@ LREC 2008*, pages 50–53. European Language Resources Association (ELRA).

Ralph Elliott, John Glauert, Vince Jennings, and Richard Kennaway. 2004. An overview of the SiGML notation and SiGMLSigning software system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation: Sign Language Processing Satellite Workshop*, pages 98–104. European Language Resources Association (ELRA).

Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal access in the information society*, 6:375–391.

Michael Filhol. 2012. Combining two synchronisation methods in a linguistic model to describe sign language. In *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication: 9th International Gesture Workshop, Revised Selected Papers*, pages 194–203. Springer Publishing.

Michael Filhol and Annelies Braffort. 2012. What constraints for representing multilinearity in sign language? In *Constraint Solving and Language Processing*.

Michael Filhol, Mohamed Hadjadj, and Annick Choisier. 2014. Non-manual features: the right to indifference. In *International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Michael Filhol and John McDonald. 2022. Representation and synthesis of geometric relocations. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*. European Language Resources Association (ELRA).

Michael Filhol, John McDonald, and Rosalee Wolfe. 2017. Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system. In *Universal Access in Human–Computer Interaction. Designing Novel Interactions*, pages 27–40, Cham, Switzerland. Springer International Publishing.

Thomas Hanke. 2004. HamNoSys: representing sign language data in language resources and language processing contexts. In *Proceedings of the fourth international conference on language resources and evaluation (LREC)*, volume 4, pages 1–6. European Language Resources Association (ELRA).

Thomas Hanke. 2021. Hamnosys.

Thomas Hanke and Jakob Storz. 2008. ilex–a database tool for integrating sign language corpus linguistics and sign language lexicography. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages*, pages 64–67. European Language Resources Association (ELRA).

Lynn Hou. 2022a. LOOKing for multi-word expressions in american sign language. *Cognitive Linguistics*, 33(2):291–337.

Lynn Hou. 2022b. A usage-based proposal for argument structure of directional verbs in american sign language. *Frontiers in Psychology*, 13(808493).

Matt Huenerfauth. 2004. Spatial representation of classifier predicates for machine translation into american sign language. In *Proceedings of the fourth international conference on language resources and evaluation (LREC)*, volume 4, pages 24–31. European Language Resources Association (ELRA).

Matt Huenerfauth. 2006. *Generating american sign language classifier predicates for english-to-asl machine translation*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Trevor Johnston and Lindsay Ferrara. 2012. Lexicalization in signed languages: When is an idiom not an idiom. In *Selected papers from UK-CLA meetings*, pages 229–248. The UK Cognitive Linguistics Association.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Richard Kennaway. 2004. Experience with and requirements for a gesture description language for synthetic animation. In *Gesture-Based Communication in Human-Computer Interaction*, pages 300–311, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ryan Lepic. 2019. A usage-based alternative to "lexicalization" in sign language linguistics. *Glossa: a journal of general linguistics*, 4(1):1–30.

Ryan Lepic and Corrine Occhino. 2018. A construction morphology approach to sign language analysis. In Geert Booij, editor, *The Construction of Words: Advances in Construction Morphology*, volume 4 of *Studies in Morphology*, pages 141–172. Springer International Publishing, Cham, Switzerland.

Martin Loetzsch, Pieter Wellens, Joachim De Beule, Joris Bleys, and Remi van Trijp. 2008. The Babel2 manual. Technical Report 01-08, AI-Memo.

I Marshall and É Sáfár. 2004. Sign language generation in an ale hpsg. In *11th International Conference on Head-Driven Phrase Structure Grammar*, pages 189–201.

Ian Marshall and Éva Sáfár. 2002. Sign language generation using HPSG. In *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages: Papers*.

Rocío Martínez, Sara Siyavoshi, and Sherman Wilcox. 2019. Advances in the study of signed languages within a cognitive perspective. *Hesperia: Anuario de filología hispánica*, 2(22):29–56.

Emmanuella Martinod, Claire Danet, and Michael Filhol. 2022. Two new azee production rules refining multiplicity in french sign language. In *Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages*. European Language Resources Association (ELRA).

Emmanuella Martinod and Michael Filhol. 2024. Formal representation of interrogation in french sign language. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 235–243. European Language Resources Association (ELRA).

Irene Murtagh. 2011a. Developing a linguistically motivated avatar for irish sign language visualisation. In *2nd International Workshop on Sign Language Translation and Avatar Technology*.

Irene Murtagh. 2011b. Towards a linguistically motivated irish sign language conversational avatar. *The ITB journal*, 12(1):73–102.

Irene Murtagh. 2020. Special nature of verbs in sign languages: An rrg account of irish sign language verbs. *TEANGA the Journal of the Irish Association of Applied Linguistics*, 11:67–99.

Irene Murtagh, Víctor Ubieto Nogales, and Josep Blat. 2022. Sign language machine translation and the sign language lexicon: A linguistically informed approach. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 240–251.

Lucie Naert, Caroline Larboulette, and Sylvie Gibet. 2020. A survey on the animation of signing avatars: From sign representation to utterance synthesis. *Computers & Graphics*, 92:76–98.

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2019. A practical guide to studying emergent communication through grounded language games. In *AISB 2019 Symposium on Language Learning for Artificial Agents*, pages 1–8. AISB.

Corrine Occhino and Sherman Wilcox. 2017. Gesture or sign? a categorization problem. *Behavioral and Brain Sciences*, 40(e66):36–37.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago, IL, USA.

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, Jan Henning, et al. 1987. HamNoSys. Hamburg Notation System for sign language. an introduction. *Zentrum für Deutsche Gebärdensprache*.

Éva Sáfár and John Glauert. 2010. Sign language HPSG. In *sign-lang@ LREC 2010*, pages 204–207. European Language Resources Association (ELRA).

Adam Schembri, Kearsy Cormier, and Jordan Fenlon. 2018. Indicating verbs as typologically unique constructions: Reconsidering verb 'agreement' in sign languages. *Glossa: a journal of general linguistics*, 3(1):1–40.

Robert Smith. 2013. Hamnosys 4.0 user guide. Technical report, Institute of Technology Blanchardstown, Blanchardstown, Ireland.

Luc Steels. 2011. Introducing Fluid Construction Grammar. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*, pages 3–30. John Benjamins, Amsterdam, Netherlands.

William C. Stokoe. 1960. Sign language structure: An outline of the visual communication system of the american deaf. *Studies in Linguistics: Occasional Papers*, (8).

Valerie Sutton. 1995. *Lessons in SignWriting: textbook & workbook*. Deaf Action Committee for Sign Writing, La Jolla, CA.

Paul Van Eecke and Katrien Beuls. 2017. Meta-layer problem solving for computational construction grammar. In *The 2017 AAAI Spring Symposium Series*, pages 258–265, Washington, D.C, USA. AAAI Press.

Paul Van Eecke, Jens Nevens, and Katrien Beuls. 2022. Neural heuristics for scaling constructional language processing. *Journal of Language Modelling*, 10(2):287–314.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

Remi van Trijp. 2015. Towards bidirectional processing models of sign language: A constructional approach in fluid construction grammar. In *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science*, pages 668–673. University of Torino.

Remi van Trijp, Katrien Beuls, and Paul Van Eecke. 2022. The FCG Editor: An innovative environment for engineering computational construction grammars. *PLOS ONE*, 17(6):e0269708.

Robert D. Van Valin Jr. 1992. *Advances in Role and Reference Grammar*. Current Issues in Linguistic Theory. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Robert D. Van Valin Jr. and William A. Foley. 1980. Role and reference grammar. In Edith Moravcsik and Jessica Wirth, editors, *Current Approaches to Syntax*, volume 13 of *Syntax and Semantics*, pages 329–352. Brill, Leiden, The Netherlands.

Sherman Wilcox and Rocío Martínez. 2020. The conceptualization of space: Places in signed language discourse. *Frontiers in Psychology*, 11(1406).

Sherman Wilcox and Rocío Martínez. 2025. Constructional approaches to signed language. In Mirjam Fried and Nikiforidou Kiki, editors, *The Cambridge Handbook of Construction Grammar*, pages 405–436. Cambridge University Press, Cambridge, United Kingdom.

Sherman Wilcox, Rocío Martínez, and Diego Morales. 2022. The conceptualization of space in signed languages: Placing the signer in narratives. In Andreas H. Jucker and Heiko Hausendorf, editors, *Pragmatics of Space*, pages 63–94. De Gruyter Mouton, Berlin, Boston.

Sherman Wilcox and Corrine Occhino. 2016. Constructing signs: Place as a symbolic structure in signed languages. *Cognitive Linguistics*, 27(3):371–404.

Sherman Wilcox and André Nogueira Xavier. 2013. A framework for unifying spoken language, signed language, and gesture. *Todas as Letras-Revista de Língua e Literatura*, 15(1):88–110.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

William A. Woods. 1968. Procedural semantics for a question-answering machine. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, pages 457–471, New York, NY, USA.

William A. Woods, Ronald M. Kaplan, and Bonnie L. Webber. 1972. The lunar sciences natural language information system: Final report. Technical report, BBN Report.

LLMs Learn Constructions That Humans Do Not Know

Jonathan Dunn[1] **and Mai Mohamed Eida**[2]
Department of Linguistics
University of Illinois Urbana-Champaign
[1]jedunn@illinois.edu
[2]maimm2@illinois.edu

Abstract

This paper investigates false positive constructions: grammatical structures which an LLM hallucinates as distinct constructions but which human introspection does not support. Both a behavioural probing task using contextual embeddings and a meta-linguistic probing task using prompts are included, allowing us to distinguish between implicit and explicit linguistic knowledge. Both methods reveal that models do indeed hallucinate constructions. We then simulate hypothesis testing to determine what would have happened if a linguist had falsely hypothesized that these hallucinated constructions do exist. The high accuracy obtained shows that such false hypotheses would have been overwhelmingly confirmed. This suggests that construction probing methods suffer from a confirmation bias and raises the issue of what unknown and incorrect syntactic knowledge these models also possess.

1 False Positives and Confirmation Bias

Recent work in computational syntax has focused on the question of whether LLMs are aware of specific syntactic structures like the LET-ALONE construction (Bonial and Tayyar Madabushi, 2024). The goal of such work is partly to evaluate the linguistic knowledge of the models themselves but also to evaluate the learnability of these constructions without specific linguistic resources available during training. Thus, constructions which an LLM does successfully learn provide evidence for learnability, especially when these constructions are relatively rare (Misra and Mahowald, 2024).

Most previous work has followed the same high-level procedure: First, a linguist relies on their own introspection to determine that a construction exists (for them) and then annotates examples of that construction in a corpus or creates examples using their own intuitions.[1] Second, these annotated examples provide stimuli for probing the linguistic knowledge of an LLM to determine whether the model is able to distinguish between this construction and other similar constructions. The procedure, in short, begins with specific constructions of interest and is limited to those constructions already hypothesized by linguists to exist for humans. For example, we start by assuming that the AANN construction exists for humans, as shown in the contrast between canonical order in (a) and non-canonical order in (b) below. Then we try to determine whether a model has also learned that construction (e.g., Mahowald 2023).

(a) five terrible weeks
(b) a terrible five weeks

If the LLM is unable to distinguish or identify the construction of interest, then the conclusion is that the model is wrong in that it disagrees with the gold-standard of human introspection (Weissweiler et al., 2022). If, on the other hand, the LLM is in fact able to distinguish this construction, this is taken as evidence that the construction is learnable from usage alone (Misra and Mahowald, 2024). At no point is it possible for the introspection-driven hypothesis that a construction exists able to be disproven. And at no point is it possible to discover that the model has also incorrectly learned other constructions that humans do not know.

There are two potential issues with this line of argumentation: First, these methods are not able to discover false positives: what constructions has an LLM learned in error? Is a model *aware of* constructions which humans do not know? In other words, by starting with constructions derived from introspection, these methods can only confirm or

[1]The exception to this is Tayyar Madabushi et al. 2020, which instead used a falsifiable if imperfect constructicon.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 13–23

disconfirm whether that specific construction has been learned. This means that it is impossible to discover that the model has learned a new construction, where such a *new construction* could be either a false positive or a construction that remains unknown to linguists or a construction from a dialect or register which linguists have yet to describe.

This paper focuses on this problem of probing for new or false positive constructions by analyzing contextual embeddings representing sets of sentences instantiating a single construction. The goal is to find constructions which the model distinguishes as separate but which humans do not.

The second potential issue with previous approaches to construction probing is that there is often no reproducible criteria to define what actually constitutes a construction from a linguistic perspective. And yet before claiming that LLMs are in error by not knowing a construction, we would want a fully reproducible and falsifiable definition of whether some pattern does in fact constitute a construction (Cappelle, 2024). How can we establish, beyond personal introspection, which constructions should be known to an LLM?

This is especially problematic considering that usage-based approaches to Construction Grammar rely on a notion of *entrenchment* in which some representations are more grammaticalized than others (Divjak, 2019). What level of entrenchment qualifies a construction as needing to be learned by an LLM and what population/register provides the baseline for measuring such entrenchment? The basic challenge is that even probing-based methods require a reproducible and falsifiable definition of what is or is not a construction, leading to a confirmation bias. Such methods can only discover that the original analysis was correct (true positives) or that the LLM is incorrect (false negatives).

The main contribution of this paper is to ask whether LLMs make errors by over-learning constructions which do not actually exist for human speakers.[2] We combine a behavioural probing task based on contextual embeddings with a meta-linguistic probing task based on prompted linguistic analysis in order to compare such false positive constructions from the perspective of both implicit and explicit linguistic knowledge.

Previous work has probed for the existence of constructions within LLMs using a variety of methods. Prompt-based approaches often rely on explicit meta-linguistic knowledge, such as asking for grammaticality judgments (Mahowald, 2023) or providing explicit descriptions or examples of constructions (Torrent et al., 2024; Bonial and Tayyar Madabushi, 2024; Morin and Larsson, 2025). Such meta-linguistic tasks require knowledge of the language but also knowledge of linguistics; many native speakers of English, for instance, would struggle with such tasks. The experiment in Section 3 uses this line of work to search for false positive constructions in explicit linguistic knowledge.

Other work has probed for constructions using more direct properties of models: log probabilities (Hawkins et al., 2020; Leong and Linzen, 2023) and contextual embeddings (Li et al., 2022; Weissweiler et al., 2022; Chronis et al., 2023). Even if a model distinguishes between constructions with similar forms, however, this does not entail that the model is able to correctly interpret that difference (Zhou et al., 2024). We refer to these more direct tasks as behavioural probes in the sense that they do not require explicit linguistic analysis in the way that the prompt-based methods do. The experiment in Section 4 uses this line of work to again search for false positive constructions, this time in implicit linguistic knowledge.

The paper is organized as follows: First, we discuss the corpus data used to probe for false positives; this consists of 100 sentences each for five clause-level constructions. Importantly, the sentences in each category are all examples of the same construction. Second, we use meta-linguistic prompts to see whether a model can be induced to mimic false analyses in a sentence sorting task. Third, we use contextual embeddings together with unsupervised methods to see whether a model distinguishes incorrectly between instances of a single construction. Together, these experiments show that LLMs hallucinate non-existing constructions and that probing experiments using these hallucinated constructions would have confirmed their existence, a significant error. These results show the need for caution in making conclusions about linguistic theory based on probing experiments.

2 Data

This section discusses the corpus data used to probe for false positive constructions. The basic idea is to collect 100 examples each for five separate clause-level constructions. Each of these examples

[2] Supplementary material is available at https://doi.org/10.17605/OSF.IO/W2XYB

should be comparable in terms of its constructional analysis, although varying in other structural and lexical and topical attributes. These sentences are collected from the Universal Dependencies English corpora (Nivre et al., 2020), chosen so that the dependency annotations can be searched for sentences which share the same form. The extracted examples are then analyzed using introspection to ensure that each set represents one and only one clause-level construction. This provides an overall corpus of 500 sentences divided into five constructional categories as described below.

The first category is derived from intransitive constructions, as shown in (1) and (2). Although the selection criteria is focused on clauses without arguments present, most of these examples contain motion-event verbs.

(1) One little boy stands up.
(2) I literally just pretty much woke up and left this morning.

The second category is derived from transitive constructions, as show in (3) and (4). The selection criteria is a clause with a single argument in which that argument is a noun phrase.

(3) Olivia played records with the living-room windows wide open.
(4) They just built a hotel in Syria.

The third category is closely related, containing transitive constructions in which the argument is an embedded clause. Examples are shown in (5) and (6). The sentences in this category have similar structures in the embedded clauses, with some degree of natural variation across them.

(5) The Great Powers realized they had to change their decision.
(6) Quinn realized that he should be going.

The fourth category contains single-argument clauses that have been passivized. Examples are shown in (7) and (8), both containing the original agent in a by-phrase.

(7) Without a valid visa, boarding will be denied by the airline.
(8) Tropical cyclones are sustained by a form of energy called latent heat.

Finally, the fifth category is double object constructions, as shown in (9) and (10). These sentences vary by whether either of the arguments are pronominal.

(9) Silent, I give his case some thought.
(10) I faxed you the promotional on the Nimitz post office.

As shown in these examples, the data consists of sentences with the same form and the same schematic meaning at the clausal level. While there are variations across examples of each category, in terms of lexical items and sub-clausal structures, they are examples of the same underlying clausal construction given the introspections of linguists. Our question is whether LLMs view these as coherent constructions (as humans do) or whether some instances within each category are viewed as distinct constructions (thus, false positives). In both experiments it turns out that the models do posit false positive constructions within these sets. Thus, we later conduct further introspective analysis to ensure that these distinctions are not linguistically-motivated by confounding factors.

3 Experiment 1: Meta-Linguistic Prompts

Our first experiment uses meta-linguistic prompts to determine whether LLMs, in this case GPT-4, hallucinate constructions that are invisible to human speakers. The basic prompting procedure is replicated from recent work on probing GPT-4 for linguistic knowledge of constructions at different levels of abstraction (Bonial and Tayyar Madabushi, 2024).

This prompt-based approach is very similar to a sentence sorting task (Li et al., 2022): the model is given the name of a construction with an example and a set of six stimuli sentences. Three of these stimuli are actual examples of the construction and the model is asked to identify them. This is the same as sorting the sentences by syntactic similarity to the example and to each other. In the original experiment, the name of the construction is drawn from the CxG literature and the examples are constructed by a linguist. For instance, the following is a possible prompt:

> From amongst the following sentences, extract the three sentences which are

instances of the LET-ALONE construction, as exemplified by the following sentence: "None of these arguments is notably strong, let alone conclusive." Output only the three sentences in three separate lines: [*Followed by six examples to sort.*]

Because we are interested in discovering false positive constructions, we cannot use the name of existing constructions in the literature. First, these may be present in the training data and thus be known to the model. But, more importantly, we want to find constructions which linguists are not aware of and thus which have no name. To overcome this problem, we create five nonce construction names which could plausibly be used in linguistic description but which also do not suggest specific existing constructions: the Pristine Exemplar construction, the Reverted Focus construction, the Alternate Application construction, the Normalized Attribution construction, and the Entrenched Objective construction. These names have not previously been used and thus could hypothetically point to previously undescribed structures.

Our data consists of 100 sentences that are instances of five clause-level constructions. For each prompt, we randomly choose an example sentence from each category and six stimuli sentences. Importantly, these sentences are all instances of the same construction given the introspections of a linguist and thus the sorting task is prompting the model to create clusters of constructions, some of which match the fake construction name and example and others of which do not match. We have two goals here: First, to determine what would happen if we asked the model to undertake a spurious linguistic analysis and, second, to determine whether any new model-driven constructions are plausible (constituting unknown constructions) or hallucinations (constituting false positive constructions).

For each category in the data set we undertake 100 unique prompts for each of the construction names above; this allows us to examine whether these invented names influence the output of the model. Each of these prompts also draws on a unique example, so that we can also examine the influence of specific examples on the output.

We operationalize this question of false positives around the stability of the sentence sorting: do the same sentences end up being clustered together regardless of the artificial construction name and the provided example? If so, this means that the LLM is consistently making a distinction between sentences which are actually instances of the same construction. If the sorting were based on the invented construction name or on the randomly chosen example sentence, then the sorting patterns would vary along these two dimensions. But if, on the other hand, the sorting is based on an underlying hallucinated construction, then the name and the example would have no influence at all on the sorting. Thus, a high stability across these dimensions would mean that neither the name nor the example influence which patterns are ultimately discovered in this task.

An alternate way of viewing this experiment is as testing a hypothetical analysis: if we assume that there is in fact a construction with the given name (e.g., the Alternate Application construction) with the given example as a good instance, how would the model behave? In this hypothetical, some of the sentences are instances of this construction and others are not. We evaluate this hypothesis by looking at whether the same sentences are consistently sorted together, either as members of the positive or of the negative category. For instance, the transitive constructions in (11) and (12) are consistently grouped together four times with four separate exemplars. This would constitute 100% agreement in the sorting. Our puppet hypothesis would thus have reached an accuracy of 100%, confirming the validity of this hallucinated construction. The sentence in (13), on the other hand, is grouped together with (11) twice but grouped separately once. This means the model would have an accuracy of 66% for our puppet hypothesis. Low agreement here means that there is no hallucinated construction.

(11) Luckily they caught the crooks before they did one on us.
(12) They have good sushi for a good price.
(13) They can cause property damage, create a mess, and produce unpleasant smells.

The results across constructions and artificial construction names is shown in Table 1. This table considers 100 random exemplars for each cell; high consistency or accuracy within a cell means that the same sorting of sentences is reached across many exemplars. The rows in the table show the invented names. Thus, if the examples influence the sorting there would be low agreement overall and if the

	Intransitive	*Transitive (NP)*	*Transitive (C)*	*Passive*	*Double Object*
Alternate Application	92.69%	92.04%	92.84%	93.88%	92.74%
Entrenched Objective	93.93%	92.53%	91.70%	94.36%	91.25%
Normalized Attribution	94.53%	92.76%	92.20%	93.34%	93.79%
Pristine Exemplar	93.01%	93.26%	90.33%	93.18%	91.41%
Reverted Focus	93.39%	93.34%	93.64%	93.98%	93.81%

Table 1: Accuracy by Percent Correct for the Consistency of Sentence Sorting by Construction Name and Across Exemplars. High Accuracy indicates that we would have confirmed an incorrect hypothesis. **These results show that the construction name has no influence on the observed sorting behaviour.**

artificial name influences the sorting there would different patterns across rows.

These results show that the sorting of sentences into two constructions is remarkably robust across both the specific exemplar given in the prompt and the name applied to the supposed construction.[3]

In other words, if we viewed this as an actual hypothesis, that these examples represent two distinct constructions with similar forms, these results would have confirmed our hypothesis. And yet we know that this is not a real distinction: each column represents one and only one construction, given the introspection of linguists. This is strong evidence that such a methodology has a confirmation bias: we could have confirmed any constructional analysis in this way. In short, either introspection is unreliable for identifying constructions (the linguist is wrong) or the model has hallucinated a constructional distinction which does not exist.[4]

Our next question is whether these new constructions which GPT-4 reliably detects are either (i) false positive hallucinations that have no linguistic regularity or (ii) meaningful constructions which were previously missed by linguistic introspection. Since the prompts reliably produce sets of sentences which the model believes represent a single construction, we use introspection to analyze some of these model-driven distinctions. To organize the data into two separate clusters, we create a vector space which captures the co-occurrence of sentences within prompt outputs; a 0 value for instance would mean that two sentences were never paired together in a response. We then use these vectors with k-means clustering to divide the sentences into two groups.

Category	**Cluster Accuracy**
Intransitive	74.8%
Transitive (NP)	71.1%
Transitive (C)	77.0%
Passive	80.5%
Double Object	74.2%

Table 2: Accuracy by Percent Correct for Clusters Learned from the Sentence Sorting Task. High Accuracy means that a sentence only occurs in pairs with other sentences in the same cluster.

The resulting clusters make a stronger case for hallucinated constructions: the previous analysis focused on pairs of sentences that were sorted together. Here it turns out that this pairwise relationship extends all the way to indirect groups in which sentences only occur with pairs of pairs of pairs. As shown in Table 2, between 71.1% and 80.5% of sentences only occur in pairs with other sentences in the same cluster, so that these clusters explain a large portion of the sorting behaviour in this experiment. This is remarkable in that this sorting is done across many unique examples across many artificial construction names. The small-scale sentence sorting prompt produces consistent groups across many iterations, thus leaving us with these larger clusters. The next question is whether these hallucinated constructions are false positives or previously unknown structures.

(14a) All tropical cyclones are driven by high heat content waters.
(14b) As in the old days, varnish is often used as a protective film against years of dirt.

[3] Note that there are a few occasions on which GPT-4 returns "None of the provided sentences match the X construction." And a single time is only one sentence returned as a match. Thus, this kind of response is technically possible but occurs only a few times.

[4] Because the model does not recognize the name of these nonce constructions, it could have been the case that this prompt is not specific enough. To check this, we tried alternate formulations which ensured that the analysis was linguistic in nature. For example, we added these sentences to the prompts: "You are a linguist who is analyzing the grammar of sentences. A construction is a syntactic unit that maps between form and meaning..." However, these alternate formulations had no significant differences from the original prompt.

(14c) Sufaat was arrested in December 2001 upon his return to Malaysia.

An introspection-based analysis shows that there is no constructional difference between sentences in the two clusters suggested by the model. For instance, passive sentences from one cluster are given above and aligned with those from the other cluster given below. Thus, (14a) and (15a) are clearly instances of the same construction, for a human, even though they are clearly separated by GPT-4. These are examples of an hallucinated construction.

(15a) Pressure for change is driven by the wish of women to choose their own fate.
(15b) In the 21st century this book is still used as one of the basic texts in modern Structural linguistics.
(15c) His chief aide in Najaf was suddenly arrested along with 13 other members of his organization.

What does the model think these constructions look like? One clue comes from some of the very rare responses in which the sentence sorting task is not undertaken because no matches are found. As mentioned above, these occur only a handful of times. Here is one example explanation of a non-match:

> None of the sentences contain a possessive pronoun subject (like *theirs*) in a subordinate clause following a verb of cognition (like *knew*), with the subject of the subordinate clause being a reverted or pronominalized NP referencing a salient set from the discourse.

This description of the example sentence is partly nonsensical but mostly far too specific to be an actual schematic construction. This provides a clue about the nature of these hallucinated constructions: they involve too specific a description over too broad a context. For instance, recent work has shown that there is a negative relationship between the size of a model and its ability to predict human reading times (in other words, showing that models with better perplexity on a test corpus make worse predictions about surprisal: Oh and Schuler 2023). The cause of this disconnect is that the model is capable of remembering infrequent patterns within very long contexts (Oh et al., 2024). Humans learn constructions precisely because they must forget the specific details of utterances and the contexts in which they occur. Constructions are remembered so that more specific details can be forgotten.

On the other hand, these results reflect the ability of LLMs to identify and, in this case, create novel patterns; dealing with novel items is an essential part of language processing (Eisenschlos et al., 2023). The challenge here arises when this ability to create new patterns is interpreted as confirmation of the original hypothesis. This experiment would have confirmed a hypothesis that (14c) and (15c) are examples of distinct constructions.

4 Experiment 2: Behavioural Probes

Our second experiment uses contextual embeddings from the Pythia 1.4b model to determine whether the model is able to distinguish between two distinct constructions which have similar forms but different meanings. The main idea, however, is that these two constructions are not actually distinct. Thus, we are evaluating a false positive distinction and, if this embedding-based probe is successful in maintaining such a distinction, this is evidence for a confirmation bias. This experiment follows probing methods previously used to sort sentences (Li et al., 2022) and to search for the English comparative correlative construction (Weissweiler et al., 2022). The challenge here is that many previous methods for probing constructional knowledge (Weissweiler et al., 2023a) are not applicable if we are looking for constructions that we do not yet know.

We take the mean embedding for each of the five hundred sentences in the dataset, averaged across the last two layers in the model. Because we are not concerned here with the contribution of specific layers, we use this averaged representation to capture the information available toward the final layer of the model. This use of pooled sentence embeddings is chosen to replicate the methods used in previous work (Li et al., 2022).

For the sake of comparison, we include two embedding conditions: First, we use the raw embeddings, which of course capture both grammatical and non-grammatical information. These unaltered representations are called *Direct Embeddings* in Tables 3 and 4. Second, we create a grammar-focused embedding for each sentence that controls for lexical differences. This is done by also extracting the embedding for a shuffled version of the sentence,

	Intransitive	*Transitive (NP)*	*Transitive (C)*	*Passive*	*Double Object*
Direct Embeddings	0.92	0.88	0.92	0.91	0.93
Grammar-Focused	0.85	0.79	0.87	0.85	0.85

Table 3: Prediction accuracy (f-score) for distinguishing between clause-level constructions using both types of embeddings. A high accuracy validates that this method is able to distinguish between actual constructions. The value for each construction is the average f-score for distinguishing it from every other construction in a binary task.

where word order is randomized. Given the sensitivity of English grammar to word order, this has the effect of removing some syntactic information, at least that which is not recoverable from lexical items (Papadimitriou et al., 2022). We then subtract this non-grammatical embedding from the original representation to create a representation which controls for lexical or topical information. These altered embeddings are called *Grammar-Focused*.

We then conduct the analysis across both sets of embeddings to ensure that any false positive constructions are not a confound of the lexical or topical attributes of the sentences. Finally, we follow this up with an introspective analysis of the results in order to search for additional possible confounds.

Our first step is to validate that these two sets of embeddings are able to correctly distinguish between the five true positive clause-level constructions in our data set. To do this, we train a logistic regression classifier with the goal of learning to distinguish between actual constructions. A high accuracy here would mean that these representations capture the grammatical generalizations that distinguish between these five constructions. These are true positives in the sense that linguists expect the grammatical representations of these constructions to be distinct. The results are shown by embedding type and construction type in Table 3; these results are averaged across five-fold cross-validation. This level of accuracy validates that, if we were probing for actual true positive constructions, these methods would confirm the existence of those constructions for the model. Interestingly, the grammar-focused embeddings are worse at distinguishing constructions in all cases.

The next step is to develop a puppet hypothesis by trying to use these embeddings to find false positive constructions. The goal is find potential fake hypotheses that would also be confirmed by these same methods. For this we use k-means clustering to divide each set of sentences into two groups. This is a simple approach of creating a false distinction within each construction: according to the introspections of linguists, each category contains 100 examples of one and only one construction. We have divided these into two groups by clustering and then use a logistic regression classifier to test whether such a division would be confirmed as an actual constructional distinction. As before, a high accuracy means that the model confirms our analysis; the difference is that this is a decoy analysis. Note that we do not control for non-constructional factors like sentence length (Weissweiler et al., 2023b), in part because we are searching for potential constructions rather than creating a test set for a hypothesized construction: we cannot manipulate these clusters. The introspection-based analysis at the end of this section, however, shows that there are no clear confounding factors in these two sets of sentences.[5]

	Direct Embeddings	*Grammar-Focused*
Intransitive	0.99	0.99
Transitive (NP)	0.94	0.93
Transitive (C)	0.96	0.99
Passive	NA	0.99
Double Object	0.97	0.97

Table 4: Prediction accuracy by f-score for distinguishing between fake puppet constructions within each constructional categories. A high f-score means that the model hallucinates additional constructions that are not distinguished for humans. Reported numbers are the mean across 5-fold cross-validation.

The results of this experiment are shown in Table 4, across both actual constructions (rows) and the embedding conditions (columns). As before, *Grammar-Focused* embeddings have been filtered to remove lexical or topical information, thus focusing more on structure. These results consistently show that the model makes distinctions between hallucinated constructions that do not exist for humans. In fact, the clarity of the hallucinated con-

[5] A quantitative analysis shows that these clusters could not be explained by factors like sentence length alone.

structions (accuracy) is higher than of the true constructions. The only exception to this is the passive construction with direct embeddings; in this case, the clustering forms a single group and no classification probe is possible. In all other cases, this false hypothesis would have been confirmed.

It is possible, of course, that the model is more correct than human linguists in its analysis of constructions. Perhaps these new constructions discovered by the model are actually correct. Thus, we now ask: are these new constructional divisions linguistically motivated? To answer this question we undertake an introspection-based analysis of the clusters, starting with the Transitive (NP) sentences. Sentences from the model's first hallucinated construction are given in (16) and from the second hallucinated construction in (17).

(16a) You can change the color of a control.
(16b) In March 1613 he bought a gatehouse in the former Blackfriars priory.
(16c) His willful nature caused trouble throughout his life.

A comparison of these examples reveals that there is no grammatical distinction between these two sets of sentences. For instance, these sentences are paired by verb, with even the same sense of the same verb in the same clausal construction existing within both groups. And yet, if we had hypothesized that these sentences were instances of two separate constructions, the model would have confirmed our analysis with an f-score of 0.94.

(17a) You change the layout by moving the fields to predefined drop areas.
(17b) He bought a postcard of brilliant blue sea and dazzling white ruins.
(17c) They can cause property damage, create a mess, and produce unpleasant smells.

A further set of examples is given in (18) and (19), in this case representing two hallucinated constructions within the clausal argument transitive sentences. According to our introspection, these are all examples of a single construction.

(18a) I realize that some were not signed by the artist.
(18b) We all know that John Kerry served in Vietnam.
(18c) I believe he must have waited among the gorse bushes through which the path winds.

As before, the examples of each hallucinated construction are aligned by verb, with (18a) comparable to (19a) and so on. And yet these do not form actual minimal pairs: each is still an instance of the same construction. These examples show that the groupings from the model, while robust, do not form a linguistically distinct set of utterances.

(19a) Nor did she realize that he wrote popular literature.
(19b) We all know that the market share of the railways has declined in recent years.
(19c) Once I returned to pick up my car, you can believe I spent quite a bit more time standing around waiting.

This section has conducted a probing experiment using contextual embeddings, first to distinguish between actual constructions and second to search for false positive constructions learned by the model. The high accuracy which validates the true positives is comparable to the high accuracy for the false positives. We then undertook a qualitative error analysis to determine if these new constructions had a legitimate but previously unknown linguistic basis. Our conclusion is that these do constitute false positive hallucinated constructions.[6]

5 Discussion and Conclusions

The goal of this paper has been to investigate the possibility of false positive constructions in LLMs. If a question for computational syntax is whether these models are *aware of* some syntactic structure, it is important to also search for constructions which the LLM is aware of incorrectly: hallucinated constructions that exist only for the model and not for humans.

This paper has shown that previous methods are inadequate for mapping the full syntactic knowledge of language models. Since we do not know how many such hallucinated constructions exist, there is a large piece missing in our understanding of how LLMs represent grammar. We can imagine two distinct scenarios: First, suppose that a human speaker of English knows 10k constructions and

[6]Importantly, previous work which included additional NLI tasks along with the identification of constructions (Weissweiler et al., 2023b) would only have over-identified in the first task.

that an LLM knows 9k of those constructions. That would be a respectable a true positive rate of 90%. But, second, suppose that the LLM knows 9k of the 10k actual constructions, but also an additional 20k hallucinated constructions. This would be a very different story, with a false positive rate exceeding the true positive rate. The problem is that previous work has been fundamentally unable to explore the possibility of false positive constructions.

From a linguistic perspective, this means that probing experiments should not yet be taken as evidence for a given linguistic analysis. Because even incorrect analyses can be confirmed in this way, we should not accept this form of evidence as support for linguistic theory itself. This means that Construction Grammar continues to struggle with falsifiability (Cappelle, 2024). In short, probing methods require minimal pairs which assume the existence of the construction to be tested. They cannot yet provide evidence for the existence of constructions themselves.

From a computational perspective, this means that we still do not know the full linguistic knowledge within LLMs because we have only looked for what we expected to find. What we do not know is the potentially vast store of incorrect constructional representations which have also been acquired by these models. Exploring the full range of such false positives remains a challenge for future work.

Minimal Pairs Cannot Be Formulated for Unknown False Positives. Many of the core probing experiments in computational syntax are focused around minimal pairs which contrast specific phenomena: examples include active/passive alternations (Leong and Linzen, 2023), dative/ditransitive alternations (Hawkins et al., 2020), and island effects (Kobzeva et al., 2023). A paradigm that relies on minimal pairs can be relatively confident in its true positives and false negatives. For instance, this kind of stimuli could be used to prove that a model does know island constraints or that a model does not know the restrained scope of the active/passive alternation. But, because hallucinated constructions are by definition unknown, it would never be possible to construct minimal pairs until they have been discovered. Thus, unless methods are developed to thoroughly search for syntactic hallucinations, we will never actually know the full range of syntactic knowledge of a model.

Entrenchment and Exposure Are Specific to Individuals. A further challenge is that, from a usage-based perspective, constructional representations are entrenched to various degrees. This means that a construction could be partially productive and it also means that the level of productivity could vary by individual (Fonteyn and Nini, 2020; Dunn and Nini, 2021) and by speech community (Hollmann and Siewierska, 2011; Dunn, 2018). In short, usage-based theory makes claims about the grammars of specific groups who have had specific linguistic experiences. The challenge is that LLMs span speech communities and represent many different populations (Dunn et al., 2024). It is reasonable to say that speakers of American English have a given construction in a spoken register. But it is not reasonable to say that all speakers of English in all registers have that construction. Thus, another challenge is to determine what the benchmark population/register is for probing experiments. Does GPT-4 need to know all of the constructions of written American English? Of spoken Nigerian English? Of Indian English? Are these false positives actually entrenched constructions for other dialects? These are important questions to ask before we claim to understand the constructional knowledge which such models possess.

Can Computational Models Ever Tell Linguists Something They Did Not Know? Previous work in constructional probing has focused always on confirming introspection-based analyses that linguists have already undertaken about phenomena that linguists already believed to be a part of the grammar. These methods as previously formulated can never discover new phenomena, and thus are unable to tell linguists something about grammar that they did not already expect to find. At the same time, as we have seen, these methods come with a confirmation bias which would incorrectly support hypotheses that we know have no basis.

One way forward is to develop methods for mapping the syntactic knowledge of LLMs which do not assume syntactic analyses from the start: what is the grammar that a model has learned, regardless of whether that grammar matches what linguists expect to find? As in this paper, such methods would do best to combine both explicit meta-linguistic knowledge with implicit behavioural knowledge. Further, since populations and individuals differ in their grammatical knowledge, it is important for such false positive probing experiments to also account for register and dialect.

References

Claire Bonial and Harish Tayyar Madabushi. 2024. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Bert Cappelle. 2024. *Can Construction Grammar Be Proven Wrong?* Elements in Construction Grammar. Cambridge University Press.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 242–261, Toronto, Canada. Association for Computational Linguistics.

Dagmar Divjak. 2019. *Frequency in Language: Memory, Attention and Learning*. Cambridge University Press.

J. Dunn. 2018. Finding variants for construction-based dialectometry: A corpus-based approach to regional cxgs. *Cognitive Linguistics*, 29(2):275–311.

J Dunn and A Nini. 2021. Production vs Perception: The Role of Individuality in Usage-Based Grammar Induction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 149–159. Association for Computational Linguistics.

Jonathan Dunn, Benjamin Adams, and Harish Tayyar Madabushi. 2024. Pre-trained language models represent some geographic populations better than others. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12966–12976, Torino, Italia. ELRA and ICCL.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Lauren Fonteyn and Andrea Nini. 2020. Individuality in syntactic variation: An investigation of the seventeenth-century gerund alternation. *Cognitive Linguistics*, 31(2):279–308.

Robert Hawkins, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663, Online. Association for Computational Linguistics.

W Hollmann and A Siewierska. 2011. The status of frequency, schemas, and identity in cognitive sociolinguistics A case study on definite article reduction. *Cognitive Linguistics*, 22(1):25–54.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2023. Neural networks can learn patterns of island-insensitivity in Norwegian. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 175–185, Amherst, MA. Association for Computational Linguistics.

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 133–144, Amherst, MA. Association for Computational Linguistics.

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.

Kyle Mahowald. 2023. A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Cameron Morin and Matti Marttinen Larsson. 2025. Large corpora and large language models: a replicable method for automating grammatical annotation. *Linguistics Vanguard*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and

surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland. Association for Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tiago Timponi Torrent, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2024. *Copilots for Linguists: AI, Constructions, and Frames*. Elements in Construction Grammar. Cambridge University Press.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023a. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2023b. Explaining pretrained language models' understanding of linguistic structures using construction grammar. *Frontiers in Artificial Intelligence*, 6.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions are so difficult that Even large language models get them right for the wrong reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811, Torino, Italia. ELRA and ICCL.

Modeling Constructional Prototypes with Sentence-BERT

Yuri V. Yerastov
Cornerstone OnDemand, 4120 Dublin Blvd, Dublin, California
yyerastov@csod.com

Abstract

This paper applies Sentence-Bert embeddings to the analysis of three competing constructions in Canadian English: *be* perfect, predicate adjective and *have* perfect. Samples are drawn from a Canadian news media database. Constructional exemplars are vectorized and mean-pooled to create constructional centroids, from which top-ranked exemplars and cross-construction similarities are calculated. Clause type distribution and definiteness marking are also examined. The embeddings-based analysis is cross-validated by a traditional quantitative study, and both lines of inquiry converge on the following tendencies: (i) prevalence of embedded – and particularly adverbial – clauses in the *be* perfect and predicate adjective constructions, (ii) prevalence of matrix clauses in the *have* perfect, (iii) prevalence of definiteness marking in the direct object of the *be* perfect, and (iv) greater statistical similarities between *be* perfects and predicate adjectives. These findings support the argument that *be* perfects function as topic-marking constructions within a usage-based framework.

1 Introduction

Canadian English has a *be* perfect construction, e.g. *I'm done my homework*, with the range of participles varying by dialect, but normally restricted to *done, finished* and occasionally *started*. The construction is similar in form and function to the *have* perfect, e.g. *I've done my homework*, as well as the predicate adjective construction, e.g. *I'm done with my homework*. For the sake of terminological clarity, the constructions in question are exemplified in Table 1; their abbreviations, listed parenthetically in the first column, are used throughout the paper for brevity.

While superficially similar, these constructions differ systematically in semantics, syntax, and discourse function. These differences are examined in this study, based on samples collected from a Canadian news media database. The analysis combines an embeddings-based approach with traditional quantitative methods. Using Sentence-BERT (SBERT) embeddings, constructional exemplars are vectorized, mean-pooled, and aggregated into constructional centroids. Exemplar similarity to centroids provides a measure of prototypicality, while cross-construction comparisons yield a similarity matrix. Clause type distributions are then analyzed and statistically validated against these embeddings-based prototypes. The analysis is further supported by quantitative evidence from direct object marking.

The study addresses two questions: 1.) How do the *be* perfect, predicate adjective, and *have* perfect compare in terms of semantic density, clause distribution, and definiteness marking? 2.) Can embeddings-based prototypes capture constructional tendencies in ways consistent with traditional corpus analysis?

The findings converge on three points: (i) the *be* perfect patterns most closely with the predicate adjective, (ii) it contrasts sharply with the *have* perfect, which predominantly codes new information in main clauses, and (iii) its pragmatic specialization lies in topic marking. The analysis contributes both to the description of Canadian English variation and to the methodological toolkit of construction grammar, showing how embeddings can model constructional prototypes within a usage-based framework.

2 Theoretical background

2.1 Sentence embeddings and construction prototypes

Theoretical work in construction grammar has commonly relied on acceptability judgments and corpus-based statistics for argumentation. An

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 24–33

Cxn	Example
be perfect (*be*)	I am done my homework I am finished my homework I am started my homework
predicate adjective (*be-with*)	I am done with my homework I am finished with my homework I am started on my homework
have perfect (*have*)	I have done my homework I have finished my homework I have started my homework

Table 1: Examples of the three constructions analyzed

embeddings-based approach might enhance and guide traditional methods, as well as amplify statistical signal through the strengths of deep learning models.

Advances in distributional semantics have enabled us to capture linguistic meaning in vectorized representations of words, phrases, and sentences. Early approaches such as word2vec (Pennington et al., 2014), GloVe (Mikolov et al., 2013), and fastText (Mikolov et al., 2018) produced static embeddings that reflected global co-occurrence patterns. While effective for short sequences and lexical slot analysis, these methods are limited in modeling pragmatic nuance.

Transformer models – and SBERT (Reimers and Gurevych, 2019) in particular – address this limitation. This family of models is better suited for modeling discourse-level relationships and sentence-level meaning. These models work best on longer stretches of text such as a multi-clausal sentence or a sequence of sentences and perform better in capturing pragmatic relationships than do static embeddings. The specific version of SBERT used in this study for inferencing is all-mpnet-base-v2 (Song et al., 2020); it was fine-tuned by Microsoft with semantic similarity tasks on a corpus of 1 billion sentence pairs.

Sentence embeddings are particularly effective for modeling constructional prototypes because they capture a mix of semantic, syntactic, and pragmatic information within a dense vector space. Mean pooling constructional exemplars allows us to abstract away a construction's most prototypical properties and create an idealized representation that defines its central meaning. The result is constructional centroids that represent abstract prototypes relative to its member exemplars.

In order to compute sentence embeddings, this study employed an inferencing technique based on mean pooling of tokens over complete sentential spans, rather than isolated clausal domains. This choice reflects the principles of the usage-based paradigm, which posits a gradient continuum from syntax through semantics to pragmatics. For instance, Hopper and Thompson (1980) show that discourse context influences grammatical choices, and Goldberg (2005, 129-165) demonstrates how information structure can constrain syntax. Given that lexical retrieval activates a web of semantic associations, it is crucial to analyze the semantic signal extending beyond the immediate clausal domain. Because neighboring clauses can provide vital semantic associations, the broader contextual analysis enables a precise characterization of a construction's placement along a continuum of lexical specificity and schematic generality.

2.2 *be* perfect in North American English

Occurrences of the *be* perfect have been documented in Canada (Hinnell, 2012; Yerastov, 2017; Murphy, 2018) and Philadelphia (Fruehwald and Myler, 2015). These attestations have been generally restricted to aspectual participles: *done, finished, started*, although other transitive participles in the *be* perfect have been documented in Southern Atlantic states and Pennsylvania (Atwood, 1953, 26-27), in Lumbee English in the US (Wolfram, 1996), and in Bungi English in Canada (Gold, 2007).

The *be* perfect is not fully abstract. It behaves like a prefab with some fixed material, in the meaning of Bybee (2006), subject to a number of constraints. Thus, the subject slot is restricted to animate referents, the participle slot favors three items only, and the direct object slot tends to be marked for definiteness, showing sensitivity to lexical idiosyncrasy (Yerastov, 2012, 2015), and requiring exhaustivity (Hinnell, 2012, 74-77).

The *be* perfect is quite distinct from its relatives and not reducible to an elliptical or surface instantiation of any other structure. For instance, semantically, it resembles the *have* perfect when it yields resultative interpretations, e.g. *I'm done dishes* "I've finished washing the dishes". In contrast to *I'm done with dishes*, the *be* perfect cannot have a stative entailment such as "I do not want to do dishes ever again". More to the point, consider the contrast between *I'm never finished with my home-*

work on time and **I'm never finished my homework on time*, where a stative interpretation of the *be* perfect fails. Finally, in dialects that do allow *start* in the construction, it is hard to induce a stative reading on an inceptive verb; *I'm started my homework* can only be interpreted resultatively.

Yet in other environments the *be* perfect behaves like its predicate adjective relative: both constructions share stative adjectival properties. For instance, the two constructions allow extent adverbs, while the *have* perfect does not, e.g. *I'm all done (with) my chores*, c.f. **I've all done my chores* (Yerastov, 2012, 444) and *I'm all ready*. Further, the *be* perfect and predicate adjective constructions cannot accept adverbial modifiers of manner, e.g. **I am carefully done my homework* (Fruehwald and Myler, 2015), c.f. **I'm carefully done with my homework*. Semantic similarities between *be* perfects and predicate adjectives can be further seen in the fact that they both generally disallow continuitive, hot-news and experiential readings.

These stative similarities have led linguists working within the generative tradition to resolve the status of the *be* perfect to a stative passive (Fruehwald and Myler, 2015; Murphy, 2018). While the *be* perfect undeniably exhibits stative passive properties in some environments, its resultative semantics and behavior are equally apparent in others. Such functional duality does not pose theoretical problems for a usage-based approach to language, adopted here, which allows for gradience of morphosyntactic categories (Barlow, 2000).

3 Methods

3.1 Data collection

Geographically, the present study is restricted to Canada. The data used in the study originated in Canadian Newsstream (formerly Canadian Newsstand), a news media database, available through many North American academic and public libraries. This choice is motivated by the low to non-existent frequency of the *be* perfect in general linguistic corpora; as an example, Yerastov (2017) provides a review of scarce search results from the Corpus of American English, the Corpus of Historical American English, the Strathy Corpus, the Bank of Canadian English, the Scottish Corpus of Texts and Speech, and Project Gutenberg – the documented attestations in these sources, while valuable, are insufficient for statistical generalizations.

Because the *be* perfect is a low-frequency, dialectally marked construction, an exhaustive search was feasible, yielding 1719 tokens. For comparison, stratified probability samples were collected for the *have* perfect (603 tokens) and predicate adjective constructions (702 tokens). Stratification ensured balanced representation across participles and tense permutations. Post-processing of the samples led to the filtering-out of sequences that did not meet target morphosyntactic criteria (e.g., misparsed complements). The end result was the difference in sample size for the three constructions. However, the resulting samples were large enough to afford meaningful statistical generalizations.

Two exclusions from the study should be noted: 1.) the participle *started*, and 2.) stand-alone and interposed adverbial clauses. Only 6 *started* exemplars were found in the *be* perfect sample. Because the baseline for the comparison was the *be* perfect, these exemplars were excluded from the study. Only 1 interposed and 3 stand-alone adverbial clauses were found in the three samples; they were omitted in the adverbial analysis due to their scarcity.

3.2 Analytical procedure

Constructional exemplars were represented by mean-pooling of token embeddings from the last hidden layer, with inference performed over the entire sentential span. While most exemplars were complete sentences, some were truncated search engine results; however, even in these cases, constructional slots and clause status information were fully preserved. Centroids of the exemplar embeddings were then computed for the three samples, using mean pooling as well.

The constructional prototypes were modeled by computing cosine similarity scores between each constructional exemplar and their respective centroid in order to rank all members of a distribution relative to its center. To ensure a focus on the most representative data, the 10 highest-ranking exemplars from each distribution were selected for further analysis. While a more extensive analysis involving longer rank lists or clustering of the exemplars would be beneficial, it was beyond the scope of the present study due to space limitations.

4 Results

4.1 Centroid similarity

The cosine similarity matrix in Table 2 shows that the *be* perfect is more similar to the predicate adjective (0.83) than to the *have* perfect (0.75). Distributional analyses of centroid-to-exemplar scores confirmed this pattern. The statistical properties of each of the *be, be-with* and *have* score distributions are visually summarized in Figure 1. Median similarity was found to be highest for the *be* perfect ($\tilde{x} = 0.3564$), followed by the predicate adjective ($\tilde{x} = 0.3418$) and the *have* perfect ($\tilde{x} = 0.3268$). However, applying a statistical test to assess central tendency is problematic in this case because centroid-to-exemplar similarities might violate the assumption of intra-sample independence: a centroid is defined by all vectors in a set.

As an alternative, approximate independence was achieved by summarizing per-exemplar similarities. To re-assess differences in the semantic density of each sample, cosine similarities were computed between all pairs within each sample. For each exemplar, its average similarity was calculated relative to all other exemplars in the same sample. These per-exemplar averages were then treated as approximately independent observations. In order to select an appropriate test of central tendency for these observations, their intra-sample normality was evaluated using the Shapiro-Wilk test. Because the *be* sample was found to deviate from normality ($p = 0.002$), the non-parametric Kruskas-Wallis test was applied to compare the medians across the three samples. The test indicated a statistically significant difference among the samples, $H(2) = 176.26, p < 0.001$; post-hoc pairwise comparisons were performed with Dunn's test using the Holm correction – all pairwise differences remained significant after adjustment ($p < 0.001$). The *be* sample exhibited the highest median of intra-sample similarity means ($\tilde{\tilde{x}} = 0.1257$), followed by *be-with* ($\tilde{\tilde{x}} = 0.1165$) and *have* ($\tilde{\tilde{x}} = 0.103$) – the same ranking as was observed in the centroid-based analysis.

The differences in the medians of cosine similarity distributions are not readily explained by variations in information quantity among the samples. All three samples were tokenized using spaCy (Honnibal et al., 2020), and their tokens counted per sentence. The *be* sample was found to have a higher median token count ($\tilde{x} = 25$) than the *have* sample ($\tilde{x} = 22$), yet the *be* sample exhibited the highest centroid-to-exemplar and per-exemplar similarity medians, contrasting with the *have* sample's lowest values. More to the point, the *be* sample has the largest number of exemplars, while the *have* sample the lowest. These relationships suggest that increased token length and exemplar count do not necessarily equate to greater semantic diversity in this comparison.

	be	be-with	have
be	−−	0.8318	0.7519
be-with	0.8318	−−	0.7517
have	0.7519	0.7517	−−

Table 2: Cosine similarity scores between constructional centroids

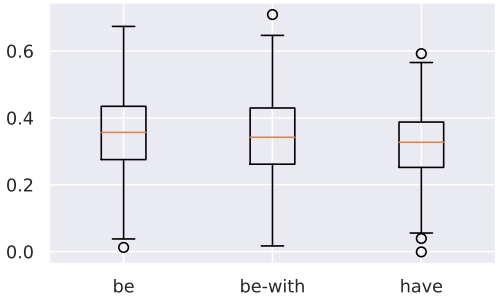


Figure 1: Distribution of centroid-to-exemplar similarity scores

The findings with respect to cosine similarity distributions allow us to evaluate the constructional samples in terms of semantic homogeneity. The more exemplars in a sample are like the center – and, more broadly, the more they are like each other – the more homogeneous the sample is overall. Therefore, it can be concluded that the *be* perfect is most semantically homogeneous, while the *have* perfect is least homogeneous (conversely, the *have* perfect is most semantically diverse); and the predicate adjective occupies a position in the middle of this continuum. From the viewpoint of construction grammar, semantic homogeneity can be interpreted as an indicator of lexical specificity, while semantic diversity – as an indicator of abstraction.

4.2 Clause type distribution

The top-ranked exemplars for the *be* perfect construction are presented in (1) through (10) sorted by cosine similarity in descending order. We observe that there are only 2 main clauses (3), (9) in

this subset, while the rest of the exemplars occur in embedded clauses. Within the embedded subset, there are 6 preposed adverbials (1), (4), (5), (6), (7), (10), 1 nominal clause (2), and 1 postposed non-finite adverbial clause of reason (8).

(1) When employees are finished that we'll send them home.

(2) I thought I'd be done school by now.

(3) By the time we got up there on Monday afternoon, they were done that part of it [...]

(4) Now my friends are done school, they're doing what they really want to do [...]

(5) Once those teachers are finished their last practicum, and they're eligible for graduation

(6) In Vancouver, when people are finished work, they're finished work.[1]

(7) When they are finished their work, they will bring it forward to us.

(8) He'll be glad to be done the homework and on to the holidays [...]

(9) I'll be done university two years from now, hopefully,

(10) Once the kids are finished school in June 1999, we'll be looking at going down.

The tendency toward embedding can also be observed in the top-ranked exemplars of the predicate adjective construction, sorted by descending similarity in (11) through (20). There are 3 preposed adverbial (13), (14), (17), 1 postposed adverbial (12), 1 relative (15), and 2 nominal (16), (18) clauses. The remaining 3 exemplars occur in main clauses (11), (19), (20).

(11) Now we are done with them.

(12) People here have made lifelong decisions because we were finished with this, Mr. Coma said.

(13) When I was finished with Mitch and Abby, I was, you know, as a creator, I was done with them, he said.

(14) When he was finished with the game, that's it, period, Gravelle said from his home in Maniwaki.

(15) It was time to have another and be done with it.

(16) If I knew I was done with this sport, it'd have been over, [Ahman Green] said.

(17) When we are finished with them, they are not finished with us.

(18) They said, for themselves, when they retired, they knew in their heart they were finished with the amateur sport world, said [Jennifer Robinson], a native of Windsor, Ont.

(19) I am done with them.

(20) We are finished here, we are done with this transaction, Einhorn, 42, told reporters on a conference call.

A distinct distributional shift is observed in the top ranked exemplars for *have* perfects, which are sorted by similarity in (21) through (30). We observe that main clauses (21), (23), (24), (25), (27), (28) have a slight edge over nominal ones (22), (26), (29), (30), with no incidence of adverbials.

(21) They have done a wonderful job and they are to be congratulated

(22) To have finished construction and started up the GTG well ahead of our schedule is an extraordinary achievement, said Derrick Kershaw, general manager of the Aurora Project.

(23) And we have done a tremendous amount of work improving our [...]

(24) This committee has done a lot of great work in the past two years [...]

(25) Our associates have done a fantastic job making sure we're ready to [...]

(26) I would have liked to have finished a little bit stronger, but to me what's important is next weekend and I'm pretty happy with today in a lot of ways, Nesbitt said in a conference call before going for a recovery massage.

(27) We have done our best and presented our best.

(28) We have finished the basic work of organizing Arts Alive in Kneehill as a registered Society in Alberta, and clarified our

[1] Here and elsewhere in the examples, when the construction of interest occurs in more than one clause within the same sentence, the more marked variant becomes the focus of analysis. Thus, this particular exemplar is treated as adverbial.

clause type	be	be-with	have
main	2	2	6
nominal	1	2	4
relative	0	1	0
adverbial	7	5	0

Table 3: Clause type distribution in top-ranked exemplars.
$G(6, n = 30) = 17.23, p = .008$

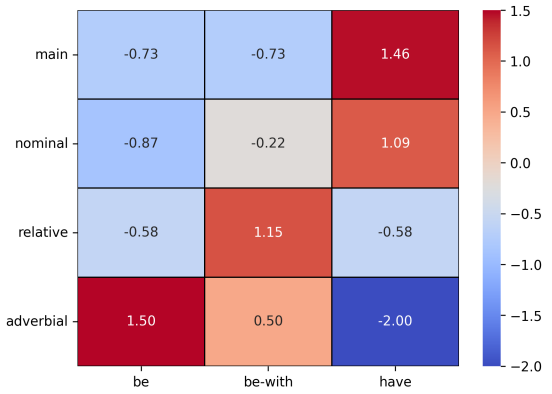


Figure 2: Standardized residuals from G-test of clause type distribution in top-ranked exemplars

mandate and goals.

(29) By now, nearly five years after he took over, it is evident that Gainey has finished the job of bringing the Canadiens out of the abyss, that awful trough where the club languished between 1999 and 2001.

(30) Hunt said council and staff have done a lot of good things over the past four [...]

To assess the distinctness of the three distributions within the centroid-based subsamples, a G-test (likelihood ratio statistic) was performed, as detailed in Table 3; this test was selected due to the limited number of observations. The statistically significant outcome ($p = .008$) supports the conclusion that the distributions differ. Further examination of the standardized residuals, shown in Figure 2, highlights a particularly strong deviation from the expected values for *have* and *be* adverbial clauses. Because this test is based on a small non-random sample, the result should be interpreted as illustrative rather than confirmatory.

5 Discussion

5.1 Semantic density

The embeddings-based prototype of the *be* perfect indicates that this construction is pragmatically rooted in recurring topical domains, particularly education and work. These patterns become especially important given that the *be* perfect displays the highest median values for both centroid-to-exemplar and per-exemplar similarity.

From a usage-based perspective – which does not draw strict boundaries between syntax, semantics, and pragmatics – such topical concentration is best interpreted as an intrinsic attribute of a construction. Specific discourse topics activate related semantic networks, thereby shaping and constraining syntactic choices. Accordingly, the informational density within a construction serves as a quantifiable measure of its lexical specificity and degree of markedness. The characteristic context of a construction is not incidental, but essential; it primes the selection of both lexical items and constructional schemas.

5.2 Adverbial clause distribution

The constructional prototype analysis reveals a strong preference for adverbial clauses among *be* perfects, and a slightly weaker adverbial tendency among predicate adjectives. In contrast, *have* perfects occur primarily in main clauses and rarely in adverbials. Since centroid exemplars represent the most prototypical instances, the absence of adverbial clauses among the top-ranked *have* perfects suggests that adverbial uses are peripheral to this construction.

The results of the prototype analysis were confirmed by a full quantitative analysis of clause type distribution across the three constructions. Table 4 presents the counts of clause types within each sample. A chi-square test of independence on these distributions revealed a statistically significant relationship between clause type and construction ($p < 0.001$), with a moderate effect size ($V = 0.294$). Based on the standardized residuals from the test, presented in Figure 3, the most extreme deviations from expectation are observed for the main *have* clauses, followed by the preposed and postposed adverbial *have* clauses – these residuals point to the *have* perfect as an outlier in the three-way comparison. Also noteworthy is the similar degree of deviation observed between the *be* and *be-with* samples with respect to main clauses.

clause type	be	be-with	have
main	437	147	408
nominal	259	141	86
postposed adv	454	141	41
preposed adv	511	254	31
relative	56	18	36

Table 4: Clause type distribution across full samples. $\chi^2(8, N = 3020) = 521.72, p < 0.001$

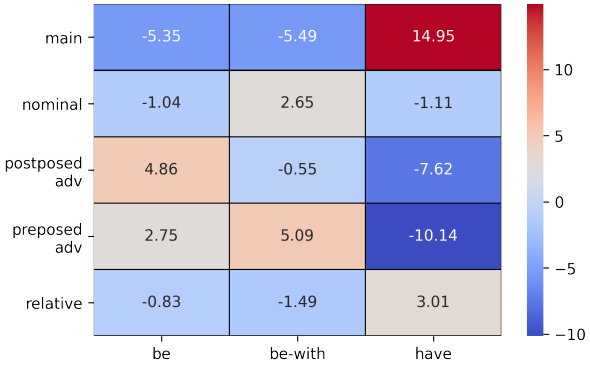


Figure 3: Standardized residuals from χ^2 test of clause type distribution

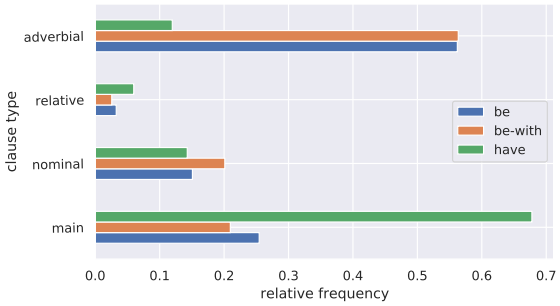


Figure 4: Distribution of constructions by clause type

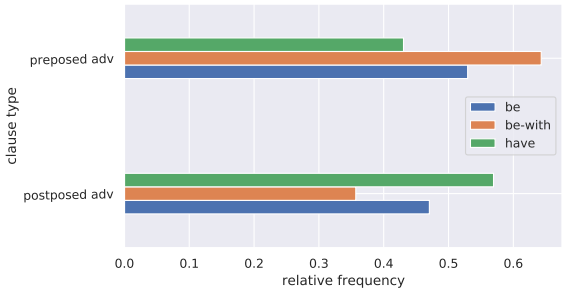


Figure 5: Distribution of constructions by adverbial clause type

To aid in the examination of the data in Table 4, the exemplar counts are normalized to relative frequencies and for better visualization presented in Figure 4 where pre- and postposed adverbials are collapsed into one class. We observe that adverbial clauses prevail in the *be* perfect ($f/n = 0.56$) and predicate adjective ($f/n = 0.56$) samples, exhibiting near identical relative frequencies, while main clauses dominate almost two-thirds of the *have* perfect sample ($f/n = 0.68$). When adverbial clauses are further isolated into separate subsamples and their counts are similarly normalized to relative frequencies (Figure 5), we find that (i) preposed adverbial clauses prevail within the *be* ($f/n = 0.53$) and *be-with* ($f/n = 0.64$) samples, and (ii) post-posed adverbial clauses prevail within the *have* sample ($f/n = 0.56$).

These quantitative findings are consistent with the pragmatic function of *have* perfects in English. It has been suggested that *have* perfects typically tend to code new information (Fenn, 1987; Michaelis, 1994; Portner, 2003); it is unusual – although not impossible – to use *have* perfects to elaborate on old information. English perfects tend to be reserved for new information, while simple pasts – for elaborations of that information (i.e. old information). The canonical – but not only – function of the *have* perfect in English is the introduction of new information. Because topic shifting is understandably less frequent in discourse than topic persistence, it is unsurprising that 68% ($n = 408$) of *have* perfect clauses in this study occur in matrix clauses rather than in embedded ones where the likelihood of presupposition and givenness is higher. And when *have* perfects do occur in adverbials they tend towards postposition. The higher incidence of postposed vis-à-vis preposed adverbial clauses among *have* perfects, revealed in this study, is in agreement with Ford (1993, 23-25), who found that the majority of perfect adverbial clauses is postposed in American English (Table 5).

While adverbial postposition is attested for both *be* and *have* perfects in this study, *be* perfect adverbials substantially outnumber *have* perfect ad-

perfect clause	count	relative freq.
postposed adv	135	0.69
preposed adv	48	0.25
stand-alone adv	11	0.06

Table 5: Distribution of present perfect by adverbial clause type in American English (adapted from Ford (1993, 24)

verbials. This outperformance is important for the present analysis because adverbial clauses are typically known to convey topical and backgrounded information (Thompson, 2011), rather than introduce new information, as would be expected from canonical perfects. Consider the exemplar in (31) from the study, in which the italicized postposed adverbial codes information of local significance, acting as a time adverbial with little anaphoric or cataphoric anchoring.

(31) To all those wonderful men who let me offer comments -- of course I looked -- and who contributed more than a loonie for the questionable privilege of me making your gift look like you'd just wrapped it yourself, thank you. [¶] The reason I was so happy -- even bursting into an off-key carol *after the elementary school kids were finished their concert*, was that each year I am heartened by the good nature of complete strangers who understand the spirit of the season is contagious. [¶] To all of those people who recognized me in the mall, bless you for reading this newspaper and helping pay my salary.

The backgrounding tendency of *be* perfect adverbials is even more evident when they are preposed; in such cases, they tend to perform global, discourse-organizing (Ramsey, 2011) functions. By way of illustration, consider two exemplars from the study. In (32), the italicized preposed adverbial follows a series of culinary descriptions and shifts topics from food to a depiction of the surrounding environment.

(32) [¶] I am a dessert lover at heart and decided to sample Ken's baklava ($4) with no regrets. This delicious dessert was crafted with several layers of phyllo pastry and walnuts. The taste of cinnamon and nutmeg were not overpowering. A clear, buttery and sweet- tasting sauce covered the entire piece. Its heat gently warmed the pastry. I could have chosen a variety of pies or muffins for dessert. [¶] *By the time I was finished my meal*, I was still quite comfortable sitting in the wooden sturdy chair at the matching table. Plenty of natural light flooded through the only large window along the front wall. A unique wall border separated the light-coloured upper wall and the lower sea-foam green coloured wall.

A similar pattern pattern can be found in (33) where the italicized preposed adverbial shifts topics from ideation to action.

(33) Before the Anti Wal-Mart War began, I had my own ideas about what the Chandler Park School could be put to use for. [¶] For years I have wanted to start a Youth Recreation Centre in Smithers, and for the past two years, I went to school to learn about business management. [¶] *Once I was finished school*, I was excited to get my plans into action - I went to Nadina and got a business plan form, and asked about small business grants. [¶] People at Community Futures Development Corporation of Nadina told me that they couldn't help me in the grant department, and gave me a form for arranging financing.

5.3 Definiteness in the direct object slot

The pragmatic specialization of the *be* perfect is evident not only in its syntactic tendencies but also in the morphology of its direct object slot. The tendency of this slot toward definiteness was already demonstrated in experimental work with native speakers of Canadian English (Yerastov, 2012, 442-443). But this study found additional, corpus-based evidence to reinforce the definiteness claim – in the context of the topicalization argument.

The direct objects of the three constructions were parsed with spaCy's (Honnibal et al., 2020) part of speech and dependency models, and slot-initial material was aggregated by category. Three cohesive categories emerged: 1) definites (definite, demonstrative and possessive determiners; demonstrative, personal and reciprocal pronouns; null anaphora); 2) indefinites (indefinite determiners and pronouns, quantifiers, wh- complementizers and relativizers); 3) undetermined nouns. Table 6 summarizes counts for each of these types across the three constructions. A chi-square test of independence revealed significant differences in these count distributions ($p < 0.001$), with a moderate effect size ($V = 0.3995$). The analysis of standardized residuals, presented in Figure 6, shows that the most pronounced deviations from expectation pertain to the indefiniteness marking of *have*

definiteness marking	be	be-with	have
definite	1290	479	230
indefinite	0	31	254
undetermined	429	192	119

Table 6: Distribution of definiteness marking in the direct object slot of the *be* perfect.
$\chi^2(4, N = 3024) = 965.41, p < 0.001$.

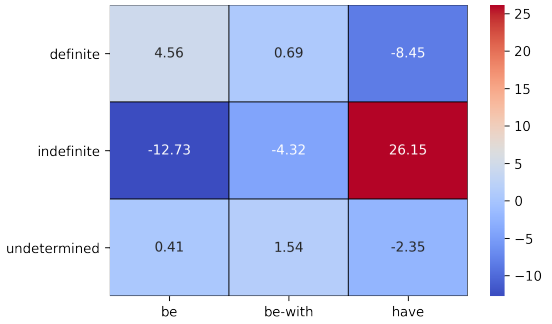


Figure 6: Standardized residuals for definiteness marking in direct object slots.

and *be* perfects.

Figure 7 visually reinforces the findings in Table 6 but in terms of relative frequency. We observe that definiteness marking of the direct object slot is strongest in the *be* perfect ($f/n = 0.75$) followed by the predicate adjective construction. Conversely, indefiniteness marking is strongest in the direct objects of *have* perfects ($f/n = 0.42$). With respect to undetermined nouns, we observe that *be-with* ($f/n = 0.27$) has a slight edge over both *be* ($f/n = 0.24$) and *have* ($f/n = 0.19$). It should be noted that undetermined noun phrases occurring in *be* perfects are either bare plurals (e.g. *chores*) or mass singulars (e.g. *school*). As such, they frequently code specific and culturally salient entities, which already carry some degree of definiteness signal in them. The prevalence of definiteness in the direct object slot is counter-expectational to the canonical tendency of direct objects in English to contain new information; it can be interpreted from a holistic perspective which takes into account the pragmatic function of the *be* perfect to background information and mark topics.

6 Conclusion

Taken together, the analyses of semantic density, clause distribution, and direct object marking converge on a unified characterization of the *be* perfect. The construction systematically patterns with predicate adjectives along the following dimensions: (i) greater semantic homogeneity, (ii) preference for embedded – and particularly adverbial – clauses, (iii) preference for pre-position among adverbial clauses, (iv) preference for definiteness marking in the direct object slot. Most importantly, the construction as a whole shows evidence of a specialized pragmatic function that consists in coding topical and backgrounded information, in clear contrast to *have* perfects, which introduce new information in matrix clauses. The topic marking function of the *be* perfect adds to the inventory of distinguishing characteristics of the construction already present in the literature (Yerastov, 2012, 2015; Hinnell, 2012; Fruehwald and Myler, 2015).

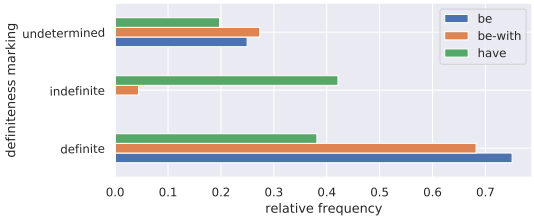


Figure 7: Relative frequency of definiteness making across the three constructions.

The distributional contrasts point to the conclusion that a constructional blend has taken place – much along the lines proposed by (Barlow, 2000), wherein syntactic, semantic, and pragmatic properties are shared across the three constructions. On the one hand, the *be* perfect inherits resultative semantics and transitive complementation from its *have* perfect relative; on the other hand, the *be* perfect inherits topicalization tendencies from its predicate adjective relative.

Methodologically, this paper demonstrates that SBERT embeddings can be used to construct prototypical representations of constructions, offering a scalable and interpretable complement to traditional quantitative analysis. The relationships observed in the constructional similarity matrix and in the centroid-based subsamples were replicated by a quantitative analysis of the entire clause type distribution. The cross-validation of these results suggests that embeddings-based methods can reliably capture distributional tendencies within a usage-based framework.

Future work should investigate the relationship between sentence-wide pragmatic signals and signals originating specifically from constructional slots.

References

E Bagby Atwood. 1953. *A Survey of Verb Forms in the Eastern United States.* University of Michigan Press.

Michael Barlow. 2000. Usage, blends and grammar. In Michael Barlow and Suzanne Kemmer, editors, *Usage-based models of language*, pages 315–345. CSLI Publications, Stanford, CA.

Joan L Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Peter Fenn. 1987. *A semantic and pragmatic examination of the English perfect*, volume 312. Gunter Narr Verlag.

Cecilia E Ford. 1993. *Grammar in interaction: Adverbial clauses in American English conversations*. Cambridge University Press.

Josef Fruehwald and Neil Myler. 2015. I'm done my homework—case assignment in a stative passive. *Linguistic Variation*, 15(2):141–168.

Elaine Gold. 2007. Aspect in bungi: Expanded progressives and be perfects. In *Proceedings of the 2007 annual conference of the Canadian Linguistic Association*, pages 1–11.

Adele E. Goldberg. 2005. *Constructions at Work: The nature of generalization in language*. Oxford University Presss.

Jennifer A. J. Hinnell. 2012. A construction analysis of [be done X] in Canadian English. Master's thesis, Simon Fraser University, Burnaby, BC.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adrienne Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Paul J. Hopper and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language*, 56(2):251–299.

Laura A Michaelis. 1994. The ambiguity of the english present perfect. *Journal of linguistics*, 30(1):111–157.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Patrick Murphy. 2018. I'm done my homework: Complement coercion and aspectual adjectives in Canadian English. *Oslo Studies in Language*, 10(2).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Paul Portner. 2003. The (temporal) semantics and (modal) pragmatics of the perfect. *Linguistics and philosophy*, 26:459–510.

Violeta Ramsey. 2011. The functional distribution of preposed and postposed 'if'and 'when'clauses in written discourse. In *Coherence and grounding in discourse: Outcome of a symposium, Eugene, Oregon, June 1984*, pages 383–408. John Benjamins Publishing Company.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Sandra A Thompson. 2011. "Subordination" and narrative event structure. In *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984*, pages 435–454. John Benjamins Publishing Company.

Walt Wolfram. 1996. Delineation and description in dialectology: The case of perfective I'm in lumbee english. *American Speech*, 71(1):5–26.

Yuri Yerastov. 2012. Transitive be perfect in Canadian English: An experimental study. *Journal of Canadian Linguistics*, 57(3):1001–1031.

Yuri Yerastov. 2015. A construction grammar analysis of the transitive be perfect in present-day Canadian English. *English Language & Linguistics*, 19(1):157–178.

Yuri Yerastov. 2017. The kids are finished school: A corpus study of geographical distribution. *Aorists and Perfects: Synchronic and diachronic perspectives*, 29:179.

Construction-Grammar Informed Parameter Efficient Fine-Tuning for Language Models

Prasanth Yadla
Independent Researcher
USA
pyadla2@alumni.ncsu.edu

Abstract

Large language models excel at statistical pattern recognition but may lack explicit understanding of constructional form-meaning correspondences that characterize human grammatical competence. This paper presents Construction-Aware LoRA (CA-LoRA), a parameter-efficient fine-tuning method that incorporates constructional templates through specialized loss functions and targeted parameter updates. We focus on five major English construction types: ditransitive, caused-motion, resultative, way-construction, and conative. Evaluation on BLiMP, CoLA, and SyntaxGym shows selective improvements: frequent patterns like ditransitive and caused-motion show improvements of approximately 3.3 and 3.5 percentage points respectively, while semi-productive constructions show minimal benefits (1.2 points). Overall performance improves by 2.4 percentage points on BLiMP and 2.4 points on SyntaxGym, while maintaining competitive performance on general NLP tasks. Our approach requires only 1.72% of trainable parameters and reduces training time by 67% compared to full fine-tuning. This work demonstrates that explicit constructional knowledge can be selectively integrated into neural language models, with effectiveness dependent on construction frequency and structural regularity.

1 Introduction

Construction Grammar fundamentally reconceptualizes linguistic knowledge as a network of form-meaning mappings called constructions, ranging from morphemes to abstract syntactic patterns (Goldberg, 1995; Fillmore et al., 1988). This theoretical framework proposes that speakers acquire grammatical competence through learning conventionalized associations between linguistic forms and their semantic interpretations, treating all linguistic knowledge as constructions of varying complexity and schematicity.

The constructionist approach offers several theoretical advantages for computational language modeling. Unlike generative approaches that separate lexicon from grammar, Construction Grammar provides a unified framework for both compositional and non-compositional linguistic phenomena. Constructions explicitly encode form-meaning correspondences, making them ideal candidates for integration into neural architectures that traditionally rely on implicit pattern recognition. The usage-based orientation of Construction Grammar aligns naturally with statistical learning paradigms underlying modern language models.

Despite these theoretical advantages, mainstream natural language processing has largely overlooked Construction Grammar insights. Current transformer-based models learn linguistic patterns through statistical exposure to large corpora but lack explicit representation of constructional knowledge (Brown et al., 2020; Devlin et al., 2018). This creates a disconnect between theoretical understanding of grammatical competence and practical implementation in language technology.

Recent work has demonstrated the potential for integrating linguistic theory into neural language models through parameter-efficient fine-tuning approaches (Hu et al., 2021). These methods enable targeted adaptation of large models while preserving general capabilities and maintaining computational efficiency. However, previous approaches have focused primarily on syntactic constraints rather than constructional form-meaning mappings.

This paper addresses this gap by introducing Construction-Aware LoRA (CA-LoRA), a parameter-efficient fine-tuning approach that explicitly integrates Construction Grammar principles into transformer-based language models. Our method treats constructions as learnable templates that specify both formal patterns and semantic interpretations, enabling models to develop explicit constructional competence.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 34–40

Benchmark	RoBERTa-large	Standard LoRA	CA-LoRA
BLiMP Overall	76.8	77.4	**79.2**
Argument Structure	73.2	74.1	**76.4**
Filler-Gap Dependencies	74.6	75.3	**77.1**
Island Effects	69.7	70.2	**71.8**
CoLA (MCC)	0.618	0.631	**0.649**
SyntaxGym	69.3	70.1	**71.7**

Table 1: Performance on linguistic evaluation benchmarks

We make four primary contributions to construction-aware language modeling. First, we develop a framework for representing major English constructions as explicit templates that can be integrated into neural training processes. Second, we present CA-LoRA, a parameter-efficient method that embeds constructional knowledge into language models through targeted parameter updates and specialized loss functions. Third, we demonstrate that constructional fine-tuning improves performance on linguistic benchmarks that test understanding of argument structure and form-meaning correspondences. Finally, we show that our approach maintains computational efficiency while achieving these linguistic competence gains.

2 Construction Grammar Framework

2.1 Theoretical Foundations

Construction Grammar emerged from recognition that traditional linguistic theories inadequately account for the pervasive role of learned form-meaning pairings in language use (Fillmore et al., 1988; Goldberg, 1995). The theory posits that linguistic knowledge consists entirely of constructions—conventionalized associations between form and meaning that speakers acquire through exposure to usage events.

Constructions exhibit several key properties that distinguish them from traditional grammatical rules. They represent holistic form-meaning mappings that cannot be derived purely through compositional processes from their component parts. They exist at multiple levels of abstraction, from fully specified lexical items to highly schematic syntactic patterns. They contribute meaning independently of their lexical fillers, explaining coercion phenomena where verbs acquire constructional semantics not present in their basic meanings.

The ditransitive construction exemplifies these principles. The pattern *[Subject Verb Object1 Object2]* carries inherent transfer semantics regardless of the specific verb involved. This explains how *"She baked him a cake"* receives a transfer interpretation despite *bake* not being inherently a transfer verb. The construction contributes transfer meaning through coercion, demonstrating how form-meaning mappings operate independently of lexical semantics.

2.2 Argument Structure Constructions

Argument structure constructions represent a well-studied domain within Construction Grammar, encompassing basic clause-level patterns that specify participant roles and event semantics (Goldberg, 1995). These constructions demonstrate clear form-meaning correspondences that extend beyond what can be predicted from lexical properties alone.

Our framework focuses on five major English argument structure constructions that exhibit systematic form-meaning relationships:

Ditransitive Construction: *[NP-Agent V NP-Recipient NP-Theme]* $\leftrightarrow$ TRANSFER(agent, theme, recipient)

This pattern encodes successful transfer events, as in *"She gave him the book"* and *"He taught her Spanish"*. The construction contributes transfer semantics that may be absent from the verb's core meaning.

Caused-Motion Construction: *[NP-Agent V NP-Theme PP-Goal]* $\leftrightarrow$ CAUSE-MOVE(agent, theme, goal)

This construction expresses caused motion events, exemplified by *"He kicked the ball into the net"* and *"She pushed the cart down the aisle"*. The pattern can coerce non-motion verbs into motion interpretations.

Resultative Construction: *[NP-Agent V NP-Patient XP-Result]* $\leftrightarrow$ CAUSE-BECOME(agent, patient, result-state)

Resultative patterns encode causation of result states, as in *"They painted the house red"* and *"He wiped the table clean"*. The construction provides result-state meaning that extends basic action se-

mantics.

Way-Construction: *[NP-Agent V Poss way PP-Path]* $\leftrightarrow$ MANNER-MOTION(agent, manner, path)

This semi-productive pattern expresses manner of motion, illustrated by *"She danced her way across the stage"* and *"He fought his way through the crowd"*. The construction creates motion interpretations for non-motion verbs.

Conative Construction: *[NP-Agent V at NP-Target]* $\leftrightarrow$ ATTEMPTED-ACTION(agent, target)

The conative alternation expresses attempted rather than successful action, contrasting *"She shot the deer"* (successful) with *"She shot at the deer"* (attempted). The prepositional marking contributes aspectual meaning.

2.3 Constructional Templates

We formalize constructions as structured templates that specify both formal constraints and semantic interpretations. Each construction C is represented as:

$$C = \langle \Phi, \Sigma, \Theta \rangle \tag{1}$$

where Φ defines the formal template including syntactic categories and linear order, Σ specifies the semantic frame with participant roles and event structure, and Θ represents frequency-based weighting derived from corpus observations.

For the ditransitive construction, this yields:

$$C_{\text{ditrans}} = \langle [\text{NP}_{\text{agent}}\ \text{V}\ \text{NP}_{\text{recipient}}\ \text{NP}_{\text{theme}}], \tag{2}$$

$$\text{TRANSFER}(\text{agent}, \text{theme}, \text{recipient}), \tag{3}$$

$$\theta_{\text{transfer}} = 0.34 \rangle \tag{4}$$

This representation captures both the syntactic pattern and associated semantic frame while incorporating usage frequency information that influences constructional processing priorities.

3 Construction-Aware LoRA

3.1 Parameter-Efficient Constructional Integration

We develop Construction-Aware LoRA (CA-LoRA), a parameter-efficient fine-tuning method that integrates constructional templates into transformer-based language models. CA-LoRA operates on the principle that constructional competence can be achieved through targeted parameter updates that encode form-meaning correspondences without disrupting general language capabilities.

The approach extends standard LoRA (Hu et al., 2021) by introducing construction-specific adaptation matrices that capture the statistical dependencies underlying each constructional pattern. For each construction C and transformer weight matrix $W_0 \in \mathbb{R}^{d \times k}$, we define construction-specific low-rank adaptations:

$$W_C = W_0 + \sum_{i=1}^{n} \alpha_i \Delta W_i^C \tag{5}$$

where $\Delta W_i^C = A_i^C B_i^C$ represents the low-rank adaptation for construction C, with $A_i^C \in \mathbb{R}^{d \times r}$ and $B_i^C \in \mathbb{R}^{r \times k}$ where $r \ll \min(d, k)$. The scaling factors α_i control the relative influence of each constructional adaptation.

This architecture allows multiple constructions to be simultaneously encoded through separate LoRA modules, enabling the model to access different constructional patterns during inference. The parameter-efficient nature ensures that constructional knowledge can be integrated without the computational overhead of full model retraining.

3.2 Construction-Guided Training Objective

We develop a specialized training objective that encourages models to learn constructional form-meaning correspondences through targeted supervision. The objective combines standard language modeling with construction-specific learning signals derived from our template representations.

The total loss function integrates multiple components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \beta \sum_{C \in \mathcal{C}} \mathcal{L}_C \tag{6}$$

where $\mathcal{L}_{\text{LM}}$ represents the standard language modeling loss, $\mathcal{C}$ denotes the construction inventory, $\mathcal{L}_C$ provides construction-specific supervision for pattern C, and β is a weighting factor that controls the relative importance of the construction losses.

Each construction-specific loss component encourages appropriate usage of the corresponding pattern:

$$\mathcal{L}_C = -\mathbb{E}_{s \sim D_C}\big[\log P(s \mid C)\big] \tag{7}$$
$$+ \lambda\, \mathbb{E}_{s \sim D_{\neg C}}\big[\max(0, \log P(s \mid C) - \tau)\big] \tag{8}$$

where D_C contains sentences that instantiate construction C, $D_{\neg C}$ contains sentences that violate constructional constraints, and τ represents a margin parameter that discourages high probability assignment to malformed patterns.

This formulation rewards models for recognizing and generating appropriate constructional patterns while penalizing violations of form-meaning correspondences. The approach enables direct supervision of constructional competence without requiring extensive manual annotation.

3.3 Multi-Construction Processing

Real language use involves interactions between multiple constructions, requiring models to handle constructional composition and selection. Our CA-LoRA framework addresses this through dynamic construction activation mechanisms that determine which patterns are relevant for specific inputs.

We implement construction selection through attention-based gating that computes relevance scores for each construction given input context:

$$w_C(x) = \text{softmax}(\text{MLP}_C(\text{pooled}(x))) \tag{9}$$

where x represents input embeddings and MLP_C provides construction-specific scoring. The final representation combines weighted contributions from all constructions:

$$h_{\text{final}} = \sum_{C \in \mathcal{C}} w_C(x) \cdot h_C(x) \tag{10}$$

This approach enables flexible constructional processing that captures the probabilistic and gradient nature of constructional activation in human language use, where multiple patterns can simultaneously influence interpretation and production.

4 Experimental Setup

4.1 Model Architecture and Training Data

We implement CA-LoRA using RoBERTa-large and GPT-2 medium as base architectures, representing both encoder-only and decoder-only transformer variants. LoRA adaptations are applied to attention projection matrices and feed-forward layers with rank $r = 16$ for attention components and $r = 32$ for feed-forward networks, based on preliminary experiments balancing expressivity with efficiency.

Training data consists of carefully selected subsets from BookCorpus (Zhu et al., 2015) and OpenWebText (Gokaslan and Cohen, 2019), totaling approximately 12GB of diverse text across multiple domains and registers. This corpus selection ensures exposure to varied constructional patterns while maintaining manageable computational requirements for parameter-efficient training.

The training process involves constructional pattern identification through template matching against our five target construction types. We use constituency parsing and semantic role labeling to identify potential constructional instantiations, then apply template matching to extract positive and negative training examples for each construction type.

Hyperparameter optimization explores construction loss weights $\beta \in \{0.1, 0.3, 0.5\}$ and margin parameters $\tau \in \{0.5, 1.0, 2.0\}$ using validation performance on a held-out subset of training data. Learning rates are tested across $\{1e{-}4, 3e{-}4, 5e{-}4\}$ with batch sizes of 16 to balance training stability with memory constraints.

4.2 Baseline Comparisons

We compare CA-LoRA against several baseline approaches that represent different methods for incorporating linguistic knowledge into language models. Standard LoRA fine-tuning provides a direct comparison, using the same training data and parameter-efficient architecture without constructional supervision.

Full fine-tuning baselines demonstrate the computational advantages of parameter-efficient approaches while providing upper bounds on potential performance gains from increased model plasticity. These models are trained on identical data with the same constructional objectives but update all model parameters.

Prompt-based approaches, while not presented here, test whether constructional knowledge can be effectively communicated through natural language descriptions rather than parameter updates, providing insights into the necessity of direct architectural integration for constructional competence.

Task	RoBERTa-large	Standard LoRA	CA-LoRA
GLUE Average	84.2	84.6	84.7
Reading Comprehension (SQuAD 2.0)	81.3	81.7	81.5
Sentiment Analysis (SST-2)	91.8	92.1	92.3
Natural Language Inference (MNLI)	86.4	86.8	86.6
Semantic Similarity (STS-B)	88.1	88.4	88.5

Table 2: Performance on general NLP tasks. Differences between Standard LoRA and CA-LoRA.

Method	Trainable Params	Training Time	Memory Usage	Performance Gain
Full Fine-tuning	355M (100%)	38.7 hours	26.8 GB	+2.1%
Standard LoRA	1.2M (0.34%)	12.4 hours	14.3 GB	+0.6%
CA-LoRA	6.1M (1.72%)	12.8 hours	14.7 GB	+2.4%

Table 3: Computational efficiency comparison for RoBERTa-large. Performance gain measured on linguistic benchmarks relative to base model.

5 Results

5.1 Linguistic Benchmark Performance

Table 1 presents results on established linguistic evaluation benchmarks, demonstrating consistent improvements from constructional fine-tuning across tasks that test grammatical competence.

CA-LoRA achieves meaningful improvements across linguistic benchmarks, with particularly notable gains of 3.2 percentage points on argument structure tasks and 2.4 points on overall BLiMP (Warstadt et al., 2020) performance. These results including CoLA (Warstadt et al., 2019) and SyntaxGym (Gauthier et al., 2020) demonstrate that explicit constructional training enhances performance on phenomena that require understanding of form-meaning correspondences and argument role relationships.

The improvements are most pronounced on tasks that directly test constructional competence, such as argument structure alternations and role assignment. This suggests that CA-LoRA successfully integrates constructional knowledge in ways that transfer to related linguistic phenomena.

5.2 Construction-Specific Analysis

Table 4 evaluates performance on tasks specifically designed to test each target construction type, providing detailed analysis of constructional learning effectiveness.

CA-LoRA demonstrates variable improvements across construction types, with gains ranging from 1.2 percentage points for way-constructions to 3.5 points for caused-motion patterns. The ditransitive construction shows a 3.3 point improvement (71.8 $\rightarrow$ 75.1), while resultative constructions show modest gains of 1.7 points (66.4 $\rightarrow$ 68.1). These results indicate that explicit constructional supervision enhances competence for well-defined form-meaning mappings, though benefits vary considerably by construction type and frequency. The performance pattern reflects both constructional frequency and structural complexity in the training data. Frequent, clearly-defined patterns like caused-motion (3.5 point improvement) and ditransitive (3.3 points) show substantial gains, while semi-productive constructions like way-constructions (1.2 points) and resultatives (1.7 points) show minimal improvement. This suggests that template-based approaches work best for constructions with clear syntactic patterns and consistent semantic roles, but struggle with more creative or contextually-dependent patterns that rely heavily on pragmatic inference.

5.3 General NLP Task Performance

Table 2 demonstrates that constructional fine-tuning maintains competitive performance on standard NLP benchmarks while achieving specialized linguistic competence.

The results show that CA-LoRA maintains performance within typical variation margins across standard benchmarks, indicating that constructional specialization does not compromise general language understanding capabilities. This supports the viability of our parameter-efficient approach for practical applications.

5.4 Computational Efficiency

Table 3 compares training costs across different approaches, highlighting the efficiency advantages of parameter-efficient constructional learning. CA-

Construction Type	Baseline	CA-LoRA
Ditransitive	71.8 ± 2.1	**75.1 ± 1.9**
Caused-Motion	69.1 ± 3.4	**72.6 ± 2.7**
Resultative	66.4 ± 2.8	68.1 ± 3.2
Way-Construction	60.2 ± 4.1	61.4 ± 3.8
Conative	64.1 ± 2.6	**67.2 ± 2.9**
Overall Average	66.3 ± 1.8	**69.0 ± 1.6**

Table 4: Construction-specific performance (accuracy %). Results averaged over 5 random seeds.

LoRA achieves superior performance gains while maintaining reasonable efficiency compared to full fine-tuning, requiring only 1.72% of trainable parameters and 67% less training time than full fine-tuning. While CA-LoRA uses approximately 5 times more parameters than standard LoRA, it remains highly parameter-efficient relative to full model retraining. The modest increase in memory usage (2.8%) reflects constructional processing overhead without fundamentally altering the parameter-efficient paradigm. The trade-off between CA-LoRA and standard LoRA involves exchanging some parameter efficiency for improved performance on linguistically-oriented tasks.

6 Analysis and Discussion

6.1 Constructional Learning Patterns

Analysis of learned parameters reveals that CA-LoRA develops distinct representational patterns for different construction types. Attention weight visualization shows increased focus on constructionally relevant features, such as recipient arguments in ditransitive constructions and result states in resultative patterns.

Probing experiments using linear classifiers demonstrate that constructional information becomes more linearly separable in CA-LoRA representations compared to baseline models. This indicates that parameter-efficient adaptation successfully embeds constructional distinctions into model representations in ways that support systematic processing.

6.2 Form-Meaning Correspondence

Qualitative analysis of model outputs demonstrates enhanced sensitivity to constructional form-meaning correspondences. CA-LoRA models show improved ability to distinguish between well-formed constructional instantiations and violations, such as correctly rejecting *"She donated him money" while accepting "She donated money to him."

The models also demonstrate better handling of constructional coercion phenomena, correctly interpreting sentences like "She sneezed the napkin off the table" where the caused-motion construction provides motion semantics absent from the verb's core meaning.

6.3 Limitations and Future Directions

Current CA-LoRA implementation focuses on English argument structure constructions and requires language-specific template definitions. Extending to other languages will need development of language-appropriate constructional inventories and consideration of typological differences in form-meaning mapping strategies.

The template-based approach may miss subtle constructional distinctions that require deeper semantic or pragmatic analysis. Future work should investigate integration of richer semantic representations and world knowledge to capture the full complexity of constructional phenomena.

Scale limitations prevent evaluation on the largest current language models, though our parameter-efficient approach should facilitate application to models with hundreds of billions of parameters. Future research should investigate how constructional learning scales with model size and training data volume.

7 Conclusion

This work demonstrates that Construction Grammar principles can be effectively integrated into neural language models through parameter-efficient fine-tuning, achieving meaningful improvements in constructional competence while maintaining computational efficiency and general language capabilities. Our Construction-Aware LoRA approach provides a practical framework for incorporating theoretical linguistic insights into modern NLP systems.

The key findings establish that explicit constructional templates can enhance language model per-

formance on tasks requiring understanding of form-meaning correspondences and argument structure relationships. Parameter-efficient methods enable integration of constructional knowledge without the computational overhead of full model retraining. Constructional fine-tuning improves linguistic competence while preserving general language understanding capabilities across diverse tasks.

Future research should explore extension to broader constructional inventories, multilingual constructional learning, and integration with larger-scale language models. Investigation of constructional learning in very large models could reveal whether explicit constructional guidance remains beneficial at scale or whether implicit statistical learning eventually captures these patterns automatically.

This work represents a step toward bridging theoretical linguistics and computational language modeling, demonstrating that Construction Grammar insights can inform and improve neural language processing systems. By explicitly encoding form-meaning correspondences, we open possibilities for more linguistically sophisticated and interpretable language models that better align with human grammatical competence.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–75.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. Accessed: 2019-09-05.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.

ASC analyzer: A Python package for measuring argument structure construction usage in English texts

Hakyung Sung[1,2] and **Kristopher Kyle**[1]
[1]Department of Linguistics, University of Oregon, Eugene, OR, USA
[2]Department of Psychology, Rochester Institute of Technology, Rochester, NY, USA
hksgla@rit.edu, kkyle2@uoregon.edu

Abstract

Argument structure constructions (ASCs) offer a theoretically grounded lens for analyzing second language (L2) proficiency, yet scalable and systematic tools for measuring their usage remain limited. This paper introduces the ASC analyzer, a publicly available Python package designed to address this gap. The analyzer automatically tags ASCs and computes 50 indices that capture diversity, proportion, frequency, and ASC-verb lemma association strength. To demonstrate its utility, we conduct both bivariate and multivariate analyses that examine the relationship between ASC-based indices and L2 writing scores.

1 Introduction

Linguistic complexity has long been recognized as an important construct in second language (L2) production research. It is commonly conceptualized along two complementary dimensions: absolute complexity and relative complexity (Bulté and Housen, 2012; Bulté et al., 2025). Absolute complexity refers to the structural properties of language, where complexity increases with the number and interrelation of constituent units. In contrast, relative complexity pertains to the cognitive effort involved in using particular forms, typically operationalized via their relative frequency and the strength of their statistical contingencies. To date, a wide range of lexicogrammatical units have been proposed to quantify complexity dimensions, including argument structure constructions (ASCs).

ASCs are clausal-level lexicogrammatical patterns, each anchored by a main verb and a specific argument configuration (e.g., Goldberg, 1995, 2013; Diessel, 2004; Ellis and Larsen-Freeman, 2009). In L2 research, two main approaches have examined their linguistic complexity. One builds on Goldberg's (1995) inheritance hierarchy, which organizes ASCs by semantic role complexity and posits that learners acquire them in a developmental sequence—from simpler constructions (e.g., simple transitives) to more complex ones (e.g., transitive resultatives). Empirical studies have operationalized this trajectory by analyzing the diversity or proportion of ASCs in learner texts (e.g., Hwang and Kim, 2023; Kim et al., 2023).

The other line of research focuses on the relationship between verbs and constructions. It posits that language learners initially tend to produce ASCs with semantically prototypical verbs (i.e., those that strongly instantiate verb-specific argument patterns), which gradually generalize into more abstract constructions (Ninio, 1999). For instance, learners may first acquire the ditransitive construction using prototypical verbs like "give" (e.g., "She **gave** him a book"), before extending it to less prototypical verbs such as "offer" or "send". This developmental trajectory has often been assessed using measures such as the relative frequency and statistical contingency between verbs and constructions (e.g., Ellis and Ferreira-Junior, 2009; Kyle and Crossley, 2017). While a growing body of empirical research has supported both developmental patterns (§ 2.1), scalable and systematic tools for extracting and analyzing ASC-based indices remain underdeveloped.

To address this gap, we present ASC analyzer, a Python package that leverages a RoBERTa-based ASC tagger (Sung and Kyle, 2024b) trained on a gold-standard ASC treebank (Sung and Kyle, 2024a). The tool automatically labels ASCs and computes a suite of indices capturing their diversity, proportion, frequency, and ASC-verb lemma association strength. We also demonstrate the application of the tool through a sample analysis of 6,482 English learner essays from the ELLIPSE corpus (Crossley et al., 2023), examining the relationship between ASC-based indices and L2 English writing proficiency.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 41–49

2 Background

2.1 Empirical findings on ASC usage in L2 production

From a usage-based constructionist perspective, language is a network of form-meaning pairings (i.e., constructions) that emerge through repeated exposure and use (Fillmore, 1988; Goldberg, 1995; Langacker, 1987). The constructions develop from actual language use and are shaped by patterns of frequency, distribution, and co-occurrence in the input and output (Bybee, 2010; Diessel, 2015; Ellis, 2012; Stefanowitsch and Gries, 2003). As learning is usage-driven, linguistic knowledge accumulates incrementally, shaped by each learner's unique language experience. Empirical studies in this framework examine constructions at varying levels of granularity (e.g., words, phrases, clauses, discourse), with particular attention to clausal-level ASCs, which are schematic form–meaning pairings that encode core semantic relations such as motion, causation, and transfer (Goldberg, 1995).

A body of L2 research has illustrated how ASC usage can be investigated across different L2 modalities and proficiency scores. In L2 writing, for example, Hwang and Kim (2023) found that more proficient L2 writers tend to produce a higher proportion of complex ASCs such as resultatives. Kim et al. (2023) further demonstrated that ASC-based indices outperform traditional T-unit measures in predicting writing proficiency. Another line of research highlights the role of verb-construction pairings. Kyle and Crossley (2017) found that L2 essay scores were negatively correlated with the relative frequency of these pairings but positively correlated with their strength of association, suggesting that advanced learners favor less frequent but more strongly associated verb–construction combinations.

Although less studied, L2 speaking shows similar patterns. Choi and Sung (2020) found that ASC use (especially transitive constructions) explained most of the variance in L2 fluency. Kim and Ro (2023) reported that advanced L2 speakers produced a wider range of verb–construction combinations. A recent study by Sung and Kyle (2025) further confirmed these findings, showing that ASC-based indices alone accounted for 44% of the variance in L2 oral proficiency scores.

2.2 ASC tagger

In this context, reliable identification of ASCs is essential for investigating their relationship with L2 proficiency in large-scale learner corpora (Kyle and Sung, 2023). To meet this need, prior studies have explored a range of tagging methodologies, including dependency parsing (e.g., O'Donnell and Ellis, 2010; Römer et al., 2014; Kyle and Crossley, 2017), rule-based approaches built on top of dependency structures (e.g., Hwang and Kim, 2023; Kim et al., 2023), and methods that leverage semantic role labels (e.g., Jeon, 2024; Kyle and Sung, 2023). Of particular relevance, Kyle and Sung (2023) introduced a supervised ASC tagger trained on a treebank that integrates semantic information across key construction types. Their system targeted nine ASC types, each defined by a characteristic mapping between semantic and syntactic frames (Appendix A).

Building on this supervised training approach and the selected ASC types, Sung and Kyle (2024b) evaluated multiple training strategies and found that a RoBERTa-based tagger trained on a combined L1 and L2 gold-standard treebank (Sung and Kyle, 2024a) achieved high F1 scores across L2 writing (0.915) and L2 speaking (0.928) domains. The results suggest that the tagger is reasonably robust across L2 production modes, providing a foundation for downstream tools that compute ASC-based indices for corpus-based L2 proficiency analysis.

3 ASC analyzer architecture

ASC analyzer is designed to compute interpretable indices that quantify the use of ASCs in English texts. Based on ASC annotations generated by the ASC tagger, the analyzer transforms these labels into a set of operational metrics.

As illustrated in Figure 1, the analyzer processes ASC-tagged output from input texts and calculates four families of indices. Diversity and proportion are text-internal measures that reflect the range and distribution of ASC types and ASC-verb lemma pair types within each text. In contrast, frequency and strength of association (SOA) are text-external measures, computed by comparing ASC usage in the input texts to norms from reference corpora, capturing how often input texts include common or strongly associated ASC-verb combinations. Note that based on the F1 scores reported in Sung and Kyle (2024b), the Gold L1+L2 model was used to

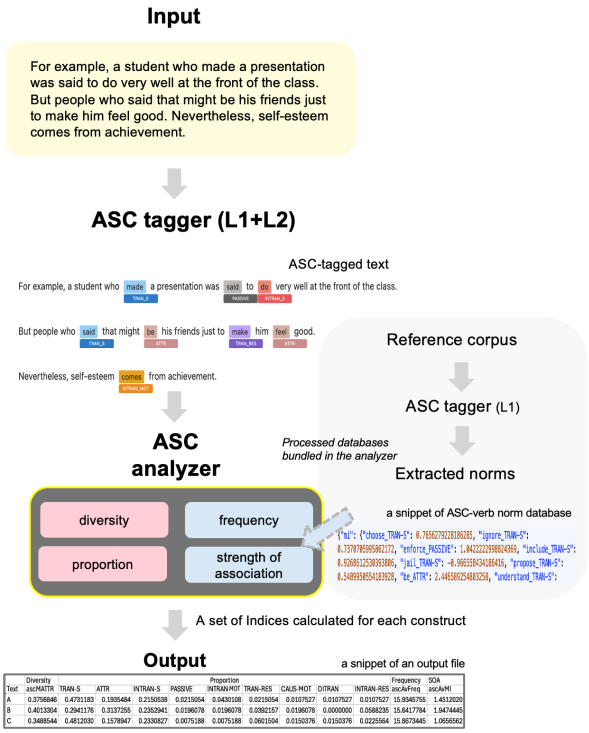


Figure 1: High-level architecture of ASC analyzer

process input texts due to its higher accuracy in L2 contexts, while the Gold L1 model was applied to the reference corpus for its more stable performance in L1 contexts. See Appendix B for detailed F1 scores of each tagger.

3.1 ASC-based Indices

Below we formalize each index family. Implementations in Python follow the equations verbatim.

Diversity. The moving-average type-token ratio (MATTR; Covington and McFall, 2010) with a sliding window w (default: 11) for a token sequence X of length N is defined as:

$$\mathrm{MATTR}_w(X) = \frac{1}{N-w+1} \sum_{i=1}^{N-w+1} \frac{|\mathrm{types}(X_{i:i+w-1})|}{w}, \quad \text{if } N \geq w+1,$$

We derive three variants: $ascMATTR$ (ASC tokens), $ascLemmaMATTR$ (ASC-verb lemma pairs), and $ascLemmaMATTRNoBe$ (ASC-verb lemma pairs excluding *be*).

Proportion. For each construction type c, we define its proportion (Hwang and Kim, 2023) in the text as

$$\mathrm{Prop}_c(X) = \frac{f_c}{N_{\mathrm{ASC}}},$$

where f_c is the number of tokens of type c in X, and N_{ASC} is the total number of ASC tokens in X. This yields nine variants, one per ASC type (e.g., $ATTR\_Prop$).

Frequency. Let an input text contain M tokens $t_1, \ldots, t_M$, where each t_i is matched up to its raw frequency $f^{\mathrm{ref}}(t_i)$ in a reference corpus (excluding types with $f^{\mathrm{ref}} < 5$). Defining

$$\ell(t_i) = \ln\big(f^{\mathrm{ref}}(t_i)\big),$$

we compute a frequency index as:

$$\mathrm{Freq} = \frac{1}{M} \sum_{i=1}^{M} \ell(t_i).$$

Two variants are derived, depending on the selected token sets: $ascAvFreq$ (ASC tokens) and $ascLemmaAvFreq$ (ASC verb-lemma pairs), following the approach of Kyle and Crossley (2017).

SOA. For each ASC–verb lemma pair (c, v), SOA scores are computed from frequency counts in a reference corpus, where $a = f_{c,v}$, $b = f_{\bar{c},v}$, $c = f_{c,\bar{v}}$, and $d = f_{\bar{c},\bar{v}}$, with total corpus size $N = a+b+c+d$. The expected frequency of the pair is given by:

$$\mathrm{E}(c,v) = \frac{(a+b)(a+c)}{N}$$

Based on these values, we define four pointwise association metrics: mutual information (MI), t-score (T), and two ΔP values:

$$\mathrm{MI}(c,v) = \log_2\left(\frac{a}{\mathrm{E}(c,v)}\right)$$

$$\mathrm{T}(c,v) = \frac{a - \mathrm{E}(c,v)}{\sqrt{a}}$$

$$\Delta\mathrm{P}_{\mathrm{Lemma}}(c,v) = \frac{a}{a+b} - \frac{c}{c+d}$$

$$\Delta\mathrm{P}_{\mathrm{Structure}}(c,v) = \frac{a}{a+c} - \frac{b}{b+d}$$

We derive two text-level SOA indices: $ascAv*$, the mean score across all ASC-lemma tokens (e.g., $ascAvMI$), and $t*$, a type-specific mean computed only over tokens labeled with ASC type t (e.g., $ATTR\_AvMI$). This indexing approach follows Gries and Ellis (2015) and Kyle and Crossley (2017).

3.2 Reference corpora for norm extraction

As briefly explained, frequency and SOA are text-external measures that reflect how closely an input text aligns with constructional norms from large external corpora. In its current version, the analyzer draws on two reference corpora:

cow We used a subset of the English Corpus of the Web (EnCOW; Schäfer, 2015; Schäfer and Bildhauer, 2012). It contains 360,783,433 tokens, 15,439,673 sentences, and 39,838,785 automatically tagged ASCs.

subt We used the SUBTLEX-US corpus of American film and television subtitles (Brysbaert et al., 2012; Brysbaert and New, 2009). The version used here comprises 76,965,430 tokens, 164,686 word types, 5,128,462 sentences, and 5,665,251 tagged ASCs across 8,388 subtitle files.

4 Using ASC analyzer: From installation to application

4.1 Installation and quick start

First, install the required dependencies and the ASC analyzer package:

```
pip install spacy
pip install spacy-transformers
python -m spacy download en_core_web_trf
pip install asc-analyzer
```

Next, view the available options:

```
python3 -m asc_analyzer.cli --help
```

To analyze a directory of input texts and save the features to CSV, run:

```
python3 -m asc_analyzer.cli \
  --input-dir "/path/to/texts" \
  --output-csv "/path/to/output.csv" \
  --source "cow"    # or "subt"
```

4.2 Application: ELLIPSE Corpus

To demonstrate the utility of the ASC analyzer in L2 research, we conducted both bivariate and multivariate analyses using a large-scale ESL writing dataset. We used 6,482 essays from the ELLIPSE corpus (Crossley et al., 2023), a reliability-filtered subset of U.S. statewide writing assessments spanning grades 8–12 across 29 prompts. Each essay includes six analytic scores for cohesion, syntax, vocabulary, phraseology, grammar, and conventions. To construct a composite proficiency index, we averaged the four subscores most aligned with constructional usage (syntax, vocabulary, phraseology, and grammar). The constructional norms were derived from the COW to compute frequency and SOA indices.

4.3 Modeling the relationship between ASC use and L2 writing proficiency

Bivariate correlations: Pearson correlations were computed between each ASC-based index and the composite writing score, retaining only those with $|r| \geq 0.10$ (Cohen, 2013).[1] As shown in Table 1, $ascMATTR$ yielded the strongest positive correlation ($r = 0.26$), while the frequency-based index $ascAvFreq$ showed the strongest negative correlation ($r = -0.22$). Although the correlations were modest overall, the results align with previous findings: more proficient L2 writers tend to use a wider variety of ASC types (Hwang and Kim, 2023) and rely less on highly frequent, but strongly entrenched, verb–construction pairings—except in the case of simple transitives (Kyle and Crossley, 2017).

Construct	Index	r
Diversity	ascMATTR	**.26**
	ascLemmaMATTR	.16
	ascLemmaMATTRNoBe	.11
Proportion	ATTR_Prop	-.11
	CAUS.MOT_Prop	.06
	DITRAN_Prop	.06
	INTRAN.MOT_Prop	.04
	INTRAN.RES_Prop	.13
	INTRAN.S_Prop	.06
	PASSIVE_Prop	.19
	TRAN.RES_Prop	.16
	TRAN.S_Prop	-.13
Frequency	ascAvFreq	**-.22**
	ascLemmaAvFreq	-.15
SOA	asc_AvMI	.12
	CAUS.MOT_AvMI	.09
	DITRAN_AvMI	.11
	INTRAN.MOT_AvMI	.08
	INTRAN.RES_AvMI	.20
	INTRAN.S_AvMI	.11
	PASSIVE_AvMI	.15
	TRAN.RES_AvMI	.14
	TRAN.S_$\Delta P_{Structure}$	-.14

Table 1: Correlations between ASC-based indices and L2 writing scores

Multivariate regression: Indices that passed the bivariate filter were entered into an AIC-based

[1] Within each SOA family, we retained only the index most strongly correlated with the scores.

model selection procedure (ΔAIC < 4; Akaike, 2003; Tan and Biswas, 2012), with multicollinearity controlled beforehand. The final model retained 12 ASC-based predictors and explained a modest proportion of variance in writing scores ($R^2_{\text{adj}} = 0.143$; Table 2), with an overall correlation of $r \approx 0.38$.

Predictor	Estimate	SE	t	p	Rel. Imp. (%)
Intercept	3.001	0.106	28.26	<.001	–
ascMATTR	0.892	0.204	4.37	<.001	16.4
ATTR_Prop	-0.902	0.106	-8.48	<.001	8.1
DITRAN_AvMI	0.013	0.003	4.58	<.001	4.2
INTRAN.RES_AvMI	0.036	0.004	9.95	<.001	15.2
INTRAN.RES_Prop	-1.082	0.502	-2.15	.031	3.4
INTRAN.S_AvMI	0.052	0.007	7.32	<.001	6.1
PASSIVE_AvMI	0.052	0.006	8.10	<.001	9.2
PASSIVE_Prop	2.242	0.294	7.62	<.001	12.2
TRAN.RES_AvMI	0.023	0.005	5.09	<.001	5.7
TRAN.RES_Prop	0.650	0.240	2.71	.007	5.9
TRAN.S_$\Delta P_{\text{Structure}}$	-8.372	1.174	-7.13	<.001	7.8
TRAN.S_Prop	-0.518	0.101	-5.14	<.001	5.7
$R^2 = 0.145$ (adj. 0.143); RSE $= 0.521$; $F(12, 6469) = 91.1$, $p < .001$					

Table 2: Summary of the regression model predicting L2 writing scores

4.4 Comparative evaluation with other models

Comparison with a syntactic complexity-based model: To evaluate the explained variance of ASC-based indices, we compared their predictive power against a multivariate model composed of syntactic complexity measures widely used in L1/L2 acquisition research (Hunt, 1965; Lu, 2011; Ortega, 2003).

Drawing from prior studies (Biber et al., 2016; Kyle, 2016; Kyle and Crossley, 2017), syntactic complexity was operationalized using a set of text-internal indices that capture structural elaboration and grammatical maturity in learner writing. These indices were grouped into three broad categories based on the syntactic units they quantify: (1) Unit length, which includes measures such as the mean length of clause; (2) Clausal complexity, which captures the frequency and depth of embedded clause constructions; and (3) Phrasal complexity, which reflects the internal modification and elaboration of noun phrases. The full list of representative indices is provided in Appendix C.

Following the same procedure used for the ASC-based indices, we built a regression model using the suite of syntactic complexity indices. The final model yielded a lower adjusted $R^2 = 0.077$. This comparison suggests that, for this corpus, ASC-based indices accounted for a greater portion of the variance in L2 writing scores than models based solely on syntactic complexity measures.

Comparison with an alternative lexicogrammatical complexity model: In studies of this kind, it is also important to examine whether newly proposed indices offer unique explanatory power beyond existing measures of lexicogrammatical complexity. To address this, we compared the ASC-based indices against a second baseline model composed of well-established lexicogrammatical indices, which primarily capture complexity at the word and bigram levels (Bulté et al., 2025).

Following prior research (Kyle and Eguchi, 2023; Paquot, 2018), the lexicogrammatical indices in this study fall into three main categories: (1) Syntactic dependency bigrams, which measure the SOA between syntactically linked words (e.g., verb and object pairs); (2) Contiguous lemmatized bigrams, which capture lexical co-occurrence patterns independent of syntactic structure; and (3) Word-level indices, reflecting lexical sophistication (e.g., frequency, concreteness, contextual and associative distinctiveness) and diversity. Full descriptions of these indices are provided in Appendix D.

The baseline model, which included only word- and phrase-level lexicogrammatical indices, achieved an adjusted $R^2 = 0.363$, while the combined model incorporating both lexicogrammatical and ASC-based indices yielded a higher adjusted $R^2 = 0.390$, reflecting an increase of $\Delta R^2 = 0.027$. This result suggests that ASC-based indices may capture an additional variance in L2 writing scores, offering complementary insights into constructional aspects of language use not fully accounted for by existing lexicogrammatical measures.

5 Conclusion

This study introduced the ASC analyzer, an open-source toolkit designed for L2 researchers and applied linguists interested in examining ASC usage in English texts. Through a proof-of-concept analysis, we demonstrated how ASC-based indices can quantify constructional patterns in L2 writing and examined their relationships with writing proficiency scores. We also compared their explanatory power against traditional syntactic complexity indices and assessed how much additional variance they capture beyond existing lexicogrammatical measures. Additional information about the package is available at `https://github.com/hksung/ASC-analyzer`.

Limitations

Several limitations should be acknowledged. First, the outputs of the ASC tagger may be influenced by model-internal biases and training data limitations, which can affect the accuracy and reliability of the extracted indices. As one reviewer noted, certain ASC types (e.g., intransitive resultatives) were underrepresented in the training data, potentially limiting performance for these low-frequency but pedagogically relevant constructions.

Second, the constructional norms used to calculate frequency and SOA scores were derived from a limited set of reference corpora. While these corpora provide useful native-speaker baselines, they may not fully capture the range of registers or genres present in the target texts.

Third, as proof of concept, this work focused on modeling rather than interpretation and did not conduct a detailed linguistic analysis of ASC usage.

Acknowledgments

This research was supported by the Harold Gulliksen Psychometric Research Fellowship (2024–2025) at ETS.

References

Hirotugu Akaike. 2003. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Douglas Biber, Bethany Gray, and Shelley Staples. 2016. Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5):639–668.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44:991–997.

Bram Bulté and Alex Housen. 2012. Defining and operationalising l2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 32:21.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, 75(2):533–574.

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

Jungyoun Choi and Min-Chang Sung. 2020. Utterance-based measurement of l2 fluency in speaking interactions: A constructionist approach. *English Teaching*, 75(1):105–126.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Holger Diessel. 2004. *The acquisition of complex sentences*, volume 105. Cambridge University Press.

Holger Diessel. 2015. Usage-based construction grammar. In Ewa Dabrowska and Dagmar Divjak, editors, *Handbook of Cognitive Linguistics*, pages 296–322. De Gruyter Mouton.

Nick C Ellis. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual review of applied linguistics*, 32:17–44.

Nick C Ellis and Fernando Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern language journal*, 93(3):370–385.

Nick C Ellis and Diane Larsen-Freeman. 2009. Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning*, 59:90–125.

Charles J Fillmore. 1988. The mechanisms of" construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

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

Adele E Goldberg. 2013. Argument structure constructions versus lexical rules or derivational verb templates. *Mind & Language*, 28(4):435–465.

Stefan Th Gries and Nick C Ellis. 2015. Statistical measures for usage-based linguistics. *Language Learning*, 65(S1):228–255.

Kellogg W Hunt. 1965. *Grammatical structures written at three grade levels*. 8. National Council of Teachers of English.

Haerim Hwang and Hyunwoo Kim. 2023. Automatic analysis of constructional diversity as a predictor of efl students' writing proficiency. *Applied Linguistics*, 44(1):127–147.

Jeeyoung Jeon. 2024. A corpus-based analysis of english argument structure constructions in esl learner writings. Master's thesis, Seoul National University. Unpublished master's thesis.

Hyunwoo Kim and Eunseok Ro. 2023. Assessment of sentence sophistication in l2 spoken production: Expansion of verbs and argument structure constructions. *System*, 119:103175.

Hyunwoo Kim, Gyu-Ho Shin, and Min-Chang Sung. 2023. Constructional complexity as a predictor of korean efl learners' writing proficiency. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, 36(2):436–466.

Kristopher Kyle. 2016. *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Ph.D. thesis, Georgia State University, Atlanta, GA.

Kristopher. Kyle and Scott Crossley. 2017. Assessing syntactic sophistication in l2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Kristopher Kyle and Masaki Eguchi. 2023. Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. *The Modern Language Journal*, 107(2):531–564.

Kristopher Kyle and Hakyung Sung. 2023. An argument structure construction treebank. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+ NLP, GURT/SyntaxFest 2023)*, pages 51–62.

Ronald W Langacker. 1987. Nouns and verbs. *Language*, pages 53–94.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL quarterly*, 45(1):36–62.

Anat Ninio. 1999. Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of child language*, 26(3):619–653.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied linguistics*, 24(4):492–518.

Matthew O'Donnell and Nick Ellis. 2010. Towards an inventory of english verb argument constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 9–16.

Magali Paquot. 2018. Phraseological competence: A missing component in university entrance language tests? insights from a study of efl learners' use of statistical collocations. *Language Assessment Quarterly*, 15(1):29–43.

Ute Römer, Matthew Brook O'Donnell, and Nick C Ellis. 2014. Second language learner knowledge of verb–argument constructions: Effects of language transfer and typology. *The Modern Language Journal*, 98(4):952–975.

Roland Schäfer. 2015. Processing and querying large web corpora with the cow14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Hakyung Sung and Kristopher Kyle. 2024a. Annotation scheme for English argument structure constructions treebank. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 12–18, St. Julians, Malta. Association for Computational Linguistics (ACL).

Hakyung Sung and Kristopher Kyle. 2024b. Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7302–7314, Miami, Florida, USA. Association for Computational Linguistics.

Hakyung Sung and Kristopher Kyle. 2025. Usage-based analysis of l2 oral proficiency: Characteristics of argument structure construction use. *Studies in Second Language Acquisition*, page 1–27.

Y.J̃. Tan, M. and Rahul Biswas. 2012. The reliability of the akaike information criterion method in cosmological model selection. *Monthly Notices of the Royal Astronomical Society*, 419(4):3292–3303.

A Target ASCs and semantic-syntactic representations

This table is reproduced from Table 1 in Sung and Kyle (2024a).

ASC (Tag)	Semantic frame	Syntactic frame
Attributive (ATTR)	theme–VERB–attribute	nsubj–cop–root
Caused-motion (CAUS_MOT)	agent–VERB–theme–destination	nsubj–root–obj–obl
Ditransitive (DITRAN)	agent–VERB–recipient–theme	nsubj–root–iobj–obj
Intransitive motion (INTRAN_MOT)	theme–VERB–goal	nsubj–root–obl
Intransitive simple (INTRAN_S)	agent–VERB	nsubj–root
Intransitive resultative (INTRAN_RES)	theme–VERB–result	nsubj–root–advmod
Passive (PASSIVE)	theme–aux–$V_{passive}$	nsubj:pass–aux:pass–root
Transitive simple (TRAN_S)	agent–VERB–theme	nsubj–root–obj
Transitive resultative (TRAN_RES)	agent–VERB–theme–result	nsubj–root–obj–xcomp

B F1 scores across ASC types by model and domain

This table, adapted from Table 2 in Sung and Kyle (2024b), reports F1 scores by ASC tag across two taggers: one trained only on the L1 treebank (Gold L1) and another trained on a combined L1+L2 treebank (Gold L1+L2). Each model is evaluated on three test sets (L1, L2 writing, and L2 speaking) to assess cross-domain robustness.

ASC Tag	Gold L1			Gold L1+L2		
	L1	L2-writing	L2-speaking	L1	L2-writing	L2-speaking
ATTR	0.972	0.954	0.986	0.968	0.971	0.988
CAUS_MOT	0.818	0.833	0.710	0.857	0.867	0.710
DITRAN	0.919	0.914	0.842	0.865	0.881	0.947
INTRAN_MOT	0.800	0.770	0.789	0.772	0.807	0.843
INTRAN_RES	0.750	0.788	0.800	0.625	0.813	0.833
INTRAN_S	0.779	0.806	0.817	0.808	0.803	0.865
PASSIVE	0.920	0.775	0.938	0.940	0.865	0.909
TRAN_RES	0.884	0.800	0.625	0.881	0.792	0.625
TRAN_S	0.931	0.929	0.927	0.936	0.943	0.948
Weighted Avg.	0.908	0.900	0.905	0.912	0.915	0.928

C Syntactic complexity indices

Dimension	Index	Description
Clause	mlc	Average number of words per finite clause
	mltu	Average number of words per T-unit
	dc_c	Number of dependent clauses per clause
	ccomp_c	Frequency of finite complement clauses
	relcl_c	Frequency of relative clauses per clause
	infinitive_prop	Proportion of "to + verb" constructions
	nonfinite_prop	Proportion of nonfinite (gerund/participial) clauses
Phrase	mean_nominal_deps	Average number of nominal dependents per noun
	relcl_nominal	Relative clauses modifying nominals
	amod_nominal	Adjectival modifiers of nominals
	det_nominal	Determiners modifying nominals
	prep_nominal	Prepositional phrases modifying nominals
	poss_nominal	Possessive modifiers of nominals
	cc_nominal	Coordinating conjunctions in noun phrases

D Word and bigram-level lexicogrammatical indices

Dimension	Index	Description
Bigram	n_amod_{T, MI, MI2, DP*}	SOA scores for noun–adjective dependencies
	v_advmod_{T, MI, MI2, DP*}	SOA scores for verb–adverb dependencies
	v_dobj_{T, MI, MI2, DP*}	SOA scores for verb–object dependencies
	v_nsubj_{T, MI, MI2, DP*}	SOA scores for verb–subject dependencies
	lemma_bg_{T, MI, MI2, DP*}	SOA scores for lemmatized word bigrams
Word	amod_freq_log	Log frequency of adjectives
	advmod_freq_log	Log frequency of adverbs
	adv_manner_freq_log	Log frequency of manner adverbs
	mverb_freq_log	Log frequency of main verbs
	lex_mverb_freq_log	Log frequency of lexical main verbs
	noun_freq_log	Log frequency of nouns
	cw_lemma_freq_log	Log frequency of content word lemmas
	b_concreteness	Word concreteness ratings
	mcd	Contextual distinctiveness (entropy-based)
	usf	Associative distinctiveness (from USF norms)
	MATTR_11	Moving-average type–token ratio (window = 11)

Note. DP* indicates various types of ΔP scores, computed using either the left or right word as the cue (or the head or dependent in the case of dependency bigrams). All scores were calculated, and for the regression model, only the score showing the strongest relationship with scores was included—consistent with the treatment of the SOA scores in the baseline model (see Footnote 1).

Verbal Predication Constructions in Universal Dependencies

William Croft
University of New Mexico
Department of Linguistics
wacroft@icloud.com

Joakim Nivre
Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

Abstract

Is the framework of Universal Dependencies (UD) compatible with findings from linguistic typology about constructions in the world's languages? To address this question, we need to systematically review how UD represents these constructions, and how it handles the range of morphosyntactic variation attested across languages. In this paper, we present the results of such a review focusing on verbal predication constructions. We find that, although UD can represent all major constructions in this area, the guidelines are not completely coherent with respect to the criteria for core argument relations and not completely systematic in the definition of subtypes for nonbasic voice constructions. To improve the overall coherence of the guidelines, we propose a number of revisions for future versions of UD.

1 Introduction

Universal Dependencies (UD) is a framework for morphosyntactic annotation, which is designed to be applicable to all human languages in a way that enables meaningful cross-linguistic comparisons (Nivre et al., 2016, 2020; de Marneffe et al., 2021). Construction grammar has also been combined with linguistic typology to allow for cross-linguistic comparison of grammatical constructions (Croft, 2016, 2022). This paper contributes to the project of adding the third edge to this triangle: representing cross-linguistically valid constructions in UD (Nivre, 2025). To find out whether UD can represent typologically justifiable constructions, Nivre (2025) proposes to build a constructicon for UD based on the survey of universal constructions and morphosyntactic realization strategies in Croft (2022) and the MoCCA database of comparative concepts derived from it (Lorenzi et al., 2024).

Croft's survey is based on two types of comparative concepts (Haspelmath, 2010; Croft, 2016): *constructions*, which are universal form-function pairings defined solely in terms of their function, and *strategies*, which are non-universal and defined by the pairing of a function with some cross-linguistically identifiable morphosyntactic form. Annotations in UD are not defined in terms of constructions and strategies, but for the framework to be universally applicable it must be possible to annotate all major constructions and strategies in the world's languages. And to support cross-linguistic comparisons, these annotations should ideally reflect systematic correspondences in constructions and strategies across languages.

The research program outlined in Nivre (2025) is to develop a constructicon for UD, consisting of the following components:

- An inventory of universal constructions.
- For each construction, an inventory of common strategies for realizing that construction in the world's languages.
- For each construction-strategy pair, a cross-linguistically valid UD analysis and representative examples from different languages.

This will help improve cross-linguistic annotation consistency by providing a complementary view of the UD guidelines, which is holistic and onomasiological; it will also provide better support for construction-based annotation on top of UD (Weissweiler et al., 2024); it will finally reveal to what extent UD can represent constructions and strategies systematically and transparently across languages, thereby identifying shortcomings in the current guidelines.

The first contribution to this project can be found in Nivre and Croft (2025) and reviews the guidelines for reference and modification constructions in UD. In this paper, we proceed to discuss verbal predication constructions, or verbal clauses, involv-

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 50–60

ing simple verbal predicates and their arguments. This family of constructions is discussed in Chapters 6–9 of Croft (2022).

2 Verbal Clause Constructions

A verbal clause construction consists of two types of elements: the head, which is a verb denoting an action or event, and argument phrases denoting participants of the action or event. This is exemplified in (1), from Croft (2022, p. 180), where the verb *broke* is associated with four argument phrases: *Sue*, *a coconut*, *for Greg*, and *with a hammer*.

(1) Sue broke a coconut for Greg with a hammer.

The grammatical encoding of argument phrases is primarily determined by their degree of salience or *topicality* to the interlocutors in the discourse. The most topical argument is encoded by the *subject*, the next most salient argument by the *object*, and all other arguments by *oblique* phrases. For example, in (1), *Susan* is the subject, *a coconut* is the object, and *for Greg* and *with a hammer* are obliques. Subjects and objects are often grouped together as *core arguments*.

In the most prototypical clause constructions, the more topical arguments are also the more central participants of the action or event. Thus, in (1), the subject denotes the agent of the action, and the object denotes the object most directly affected by the action, while the oblique arguments denote more peripheral participants. Such constructions are called *basic voice* constructions and are discussed in Section 3. In Section 4, we then turn to constructions that have been conventionalized to express non-prototypical combinations of participant roles and argument salience.

3 Basic Voice Constructions

Basic voice constructions are traditionally classified based on the number of central participant roles, or core arguments, into intransitive, transitive, and ditransitive constructions. We will begin with the transitive construction, with two core arguments, which is generally assumed to be the prototypical verbal clause (Croft, 2022, p. 183).

3.1 The Transitive Construction

If the transitive construction is the most prototypical verbal clause construction, the most prototypical event type expressed through this construction is an agentive change of state event, that is, an event where an external volitional agent brings about a change in a patient. The asymmetric semantic relation between agent and patient is force-dynamic, that is, the change of state event involves a transmission of force from the agent to the patient (Talmy, 1988; Croft, 2010). To facilitate cross-linguistic comparison, typologists have proposed that the construction be defined by an even more specific event type, the agentive breaking event exemplified in (1) (Haspelmath, 2011, 2015; Croft, 2022).

In the prototypical transitive clause, the phrase expressing the agent (A) role is the subject, and the phrase expressing the patient (P) role the object. But the same construction is commonly used also to express other event types with other semantic roles, such as motion events or experiential events. Thus, in a sentence like *she entered the cave*, the subject (*she*) expresses the figure role (F), and the object (*the cave*) expresses the ground role (G). And in *she saw the sun*, the subject (*she*) is an experiencer (X) and the object (*the sun*) is a stimulus (M).

There are cross-linguistic generalizations about the tendency for different event types to recruit[1] the transitive construction, often summarized in so-called transitivity hierarchies (Tsunoda, 1981, 1985; Malchukov, 2005; Beavers, 2011). To map out the distribution of the transitive construction in a given language, we need to study how the subject and object are encoded in prototypical transitive clauses and see to what extent the same encoding appears with other event types and semantic roles. Generally speaking, there are three common strategies used to distinguish arguments in verbal clauses, including transitive clauses, exemplified in (2–4) (Croft, 2022, pp. 187–188).

(2) Tanj-a ubi-la Mašu
Tanya-F.NOM kill-PST:FSG Masha-F.ACC
'Tanya killed Masha'

(3) x-Ø-uu-choy chee7 tza7n ikaj
PST-3SG.ABS-3SG.ERG-cut tree with axe
'he cut tree(s) with an axe'

(4) ka'se'kaw: samlap ko:n kru:k
farmer kill child pig
'(the) farmer(s) kills/killed (the) piglet(s)'

The Russian example (2), from Comrie (1989), exemplifies the use of *flags*, morphemes that encode the semantic relationship between the participant

[1] Recruitment is a relationship between two constructions in which the structure of one construction is recruited for use, or extended to use, in the other construction (Croft, in press).

and the event. In this example, flags take the form of case affixes, but flags can also be realized as adpositions. Cross-linguistically, there is a tendency for argument phrases lacking overt flags to express core argument roles.

The Tzutujil example (3), from Dayley (1989), illustrates the strategy of *indexation*, where an argument is indexed by a morpheme that is typically an affix of the predicate. In this case, both the subject and the object are indexed on the verb. Cross-linguistically, there is a strong tendency that indexed arguments express core argument roles.

Since flags occur on arguments and indexation occurs on the predicate, the two strategies may be used together. This is the case in the Russian example (2), where the subject *Tanja* carries a flag and is also indexed on the verb.

The Khmer example (4), from Haiman (2011), uses neither flags nor indexation, and the arguments are distinguished only through *word order*. The cross-linguistic study of basic word order in transitive clauses goes back to Greenberg (1966) and has shown that it is overwhelmingly more common for subjects to precede objects in languages that have a dominant order.

3.2 The Intransitive Construction

The intransitive clause construction involves a verb and a single core argument, whose role is called S by typologists, which is almost always encoded like one of the two core arguments in the transitive construction. The encoding patterns of the three roles is called *alignment* and the three most common patterns are *neutral* (A = S = P), *accusative* (A = S ≠ P) and *ergative* (A ≠ S = P) alignment. Neutral alignment is found in English when no argument is realized as a pronoun, as shown in example (5).Example (6), from Weber (1989), shows accusative alignment in Huallaga Quechua, involving both flags and indexation; example (7), from Williams (1980), shows ergative alignment with flags in Yuwaalaraay.

(5) a. the dog barked
b. the dog chased the cat

(6) a. yaku-Ø timpu-yka-n
water.NOM boil.IPFV-3
'the water is boiling'
b. Hwan-Ø Tumas-ta maka-n
John.NOM Tom.ACC hit.-3
'John hits Tom'

(7) a. wa:l nama yinar-Ø banaga-ni
NEG that woman-ABS run-NFUT
'the woman didn't run'
b. duyu-gu nama dayn-Ø yi:-y
snake-ERG that man-ABS bite-NFUT
'the snake bit the man'

Regardless of the alignment, however, the single core argument in an intransitive clause is classified as a subject, because it is the single most topical argument of the construction.

3.3 Reflexives and Reciprocals

In addition to the transitive and intransitive constructions, there are two constructions that have affinities with both and often employ the same strategies: the *reflexive* and the *reciprocal* construction. The reflexive construction expresses an event with a single participant (like intransitives) but two distinct roles (like transitives), as in *she injured herself*. The reciprocal construction expresses an event with a pair of participants that both assume the same two roles, as in *they touched each other*.

Both reflexives and reciprocals typically recruit either the transitive or the intransitive construction for their realization. The former is the *dual-role* strategy, with two argument phrases, as in the examples above, and may involve a specialized argument expression such as a reflexive or reciprocal pronoun. The latter is the *single-role* strategy, with only one argument phrase, as in *he shaved* and *they met*. Cross-linguistically, the dual-role strategy often grammaticalizes into the single-role strategy through fusion of a specialized reflexive/reciprocal element with the verb (Croft, 2022, pp. 208–209).

3.4 The Ditransitive Construction

The ditransitive clause construction is defined in terms of *transfer events*, physical transfer events expressed by verbs like *give* and *sell*, as well as mental transfer events expressed by verbs like *show* and *tell*. The roles associated with these events are agent (A), theme (T), and recipient (R). There is a force-dynamic ordering A > T > R, but R is almost always human and hence topical enough to be encoded as a core argument, comparable to T. While the A role appears to be universally encoded as the grammatical subject, languages use different strategies for encoding the T and R roles, which can again be described in terms of alignment with the transitive construction. In the *neutral* alignment (T = P = R), or double-object strategy, both

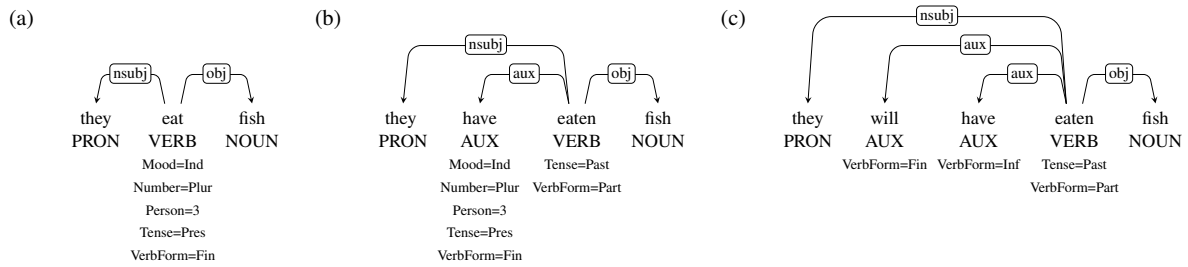


Figure 1: UD annotation of verbal predicates: (a) finite main verb, (b–c) main verb with auxiliaries.

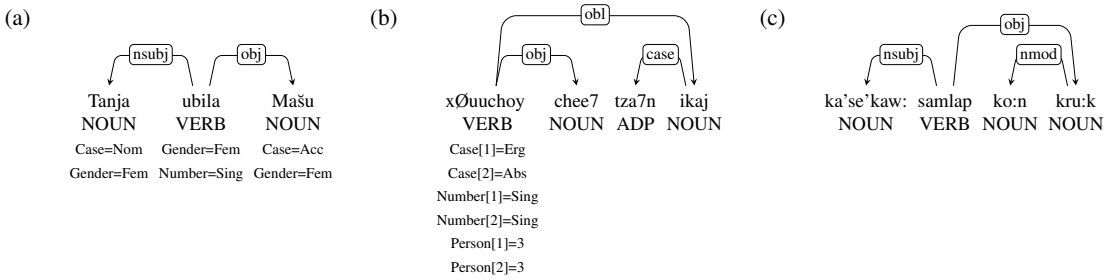


Figure 2: UD annotation of encoding strategies: (a) flags, (b) indexation, (c) word order.

T and R are co-expressed with P. In the *indirective* alignment (T = P ≠ R), T and P are co-expressed and referred to as the *direct* object, while R has a distinct encoding and is referred to as the *indirect* object. In the *secundative* alignment (T ≠ P = R), finally, R and P are co-expressed and distinct from T. In this case, the phrase expressing R or P is the *primary* object, while the phrase expressing T is the *secondary* object.

3.5 UD Annotation

When reviewing the UD annotation of basic voice constructions, our discussion will focus on how UD treats different *alignment* strategies across intransitives, transitives and ditransitives. Before we turn to that discussion, however, we will briefly review how UD annotates *verbal predicates* and how it handles the *encoding strategies* used to distinguish arguments in any of these constructions.

Verbal Predicates

The predicate of a verbal clause consists of a main verb, which is assigned the part-of-speech tag VERB, possibly together with one or more auxiliaries, which are assigned the tag AUX. Both main verbs and auxiliaries may be assigned morphological features capturing properties such as tense, mood, and aspect. It is worth noting that UD always treats the main (lexical) verb as the root of the clausal structure, regardless of whether it is finite or not, and attaches auxiliaries to the main verb with the syntactic relation *aux*, as illustrated in Figure 1.[2]

Encoding Strategies

As observed in Section 3.1, there are three main strategies used to distinguish arguments: flags, indexation, and word order. These are annotated to varying degrees in UD, using part-of-speech tags, morphological features, and relations:

- Flags realized as morphological affixes are represented by the morphological feature Case, as shown for example (2) in Figure2(a), while adpositions are tagged ADP and attached with the *case* relation, as exemplified by the oblique argument in Figure 2(b).
- Indexation is also represented by morphological features, whose values correspond to those of the indexed arguments, as shown for example (3) in Figure 2(b). When multiple arguments are indexed, as in this example, the technique known as *layering* is used to represent multiple values of the same feature. However, as observed by Nivre and Croft (2025), there

[2]In these and all following examples, we simplify the UD representations by omitting (a) lemmas and (b) morphological features that are not relevant for discussion (notably features on nominal arguments in these examples and features on verbal predicates in subsequent examples).

is nothing in the annotation that explicitly connects the index features to the arguments.

- The word order strategy is not annotated explicitly, but word order is preserved in the representation; cf. example (4) and Figure 2(c).

Intransitive–Transitive Alignment

The intransitive construction is annotated in UD by attaching the single core argument to the verb with the *nsubj* relation. This is consistent with the analysis in Croft (2022) in that the phrase expressing the S role is analyzed as the grammatical subject.

For the transitive construction, the idea is to use *nsubj* and *obj* for any two arguments encoded as the A and P arguments of a prototypical transitive clause describing an agentive change-of-state event, including clauses describing motion events (*she entered the cave*) and experiential events (*she sees the sun*). However, if one of the arguments has an oblique encoding, then it is instead annotated with the *obl* relation, even if it expresses the same role as in the corresponding transitive clause. Thus, a clause like *she ran into the cave* is analyzed as an intransitive clause, with *she* as *nsubj* and *into the cave* as *obl*, and similarly for a clause like *she looked at the sun*. Of course, oblique arguments may also appear in transitive clauses, as in caused motion events like *she chased them into the cave* (*she* = *nsubj*, *them* = *obj*, *into the cave* = *obl*).

The question, however, is how to identify subjects and objects in languages with different alignment strategies. The documentation on the UD website[3] appears to follow Croft (2022) in treating the phrase expressing the A role in a prototypical transitive clause as the grammatical subject regardless of alignment, because it is the most topical argument. More specifically, it says that "case alignment should not be used to decide the assignment of core argument roles" and that "in ergative languages, the patient-like argument of a transitive verb (O/P) will take the the *obj* relation despite the fact that it carries the same case marking as the *nsubj* argument (S) of an intransitive verb". The annotations in Figure 2 are compatible with these guidelines, specifically Figure 2(b), where the argument indexed with absolutive case is analyzed as *obj*. However, in a more detailed discussion of ergativity, de Marneffe et al. (2021) argue that, while this analysis is appropriate for languages where ergative–absolutive case marking is primarily a morphological feature, such as Basque, there are other languages, such as Jirrbal (or Dyirbal), where ergativity extends to syntactic relations. For such languages, de Marneffe et al. (2021) propose an analysis based on Dixon (1994), where the S and P arguments are treated as a "pivot" and are both assigned the *nsubj* relation, while the A argument is instead assigned the *obj* relation. To indicate the unusual role assignment, it is recommended to use the subtype *nsubj:pass*[4] for the P argument and the subtype *obj:agent* for the A argument (de Marneffe et al., 2021, p. 295).

Reflexives and Reciprocals

Reflexive and reciprocal constructions are in principle annotated exactly as the constructions they recruit, that is, the transitive or intransitive construction. However, if a language employs a specialized dual-role strategy involving a reflexive or reciprocal pronoun, this may be captured by features on the pronoun, such as Reflexive=Yes and PronType=Rcp. The UD guidelines also prescribe a special treatment of so-called inherent reflexive verbs, such as *se souvenir* (remember) in French, where the verb cannot occur with a non-reflexive pronoun and where there is arguably only one semantic role. In this case, the reflexive pronoun should be attached to the verb with the *expl* (expletive) relation (instead of the *obj* relation) to indicate that it does not express a semantic role in relation to the predicate.

Ditransitive–Transitive Alignment

UD defines ditransitive clauses more narrowly than Croft (2022) and only has specific guidelines for the neutral alignment strategy, where the T and R roles are both encoded as core arguments. In this case, UD assigns *nsubj* to the A argument, *obj* to the T argument, and a special relation *iobj* (for indirect object) to the R argument. The *obj*/*iobj* distinction is upheld even if the T and R arguments have identical encoding, and is thus based on roles rather than morphosyntactic realization.

For the *indirective* strategy, UD uses *nsubj* for the A argument, *obj* for the T argument, and *obl* for the R argument with an oblique encoding, which typically involves either an adposition, as in English *she gave the book to Peter*, or morphological

[3] https://universaldependencies.org/u/overview/simple-syntax.html#intransitive-and-transitive-clauses

[4] The subtype *:pass* was first used in the analysis of passive constructions (hence the name), but it is now used more generally in UD for subjects whose semantic role is lower than expected in the transitivity hierarchy.

Construction	Strategy	Roles					
		S	**A**	**P**	**T**	**R**	**C**
Intransitive	–	nsubj					
Transitive	Accusative		nsubj	obj			
	Ergative 1		nsubj	obj			
	Ergative 2		obj:agent	nsubj:pass			
Ditransitive	Neutral		nsubj		obj	iobj	
	Indirective		nsubj		obj	obl	
	Secundative		nsubj		obl	obj	
Construction	**Basic Voice**						
Passive	Transitive		obl:agent	nsubj:pass			
Causative	Intransitive		obj:caus				nsubj
	Transitive		iobj:caus	obj			nsubj

Table 1: UD relations for semantic roles in verbal clause constructions (C = external causer).

case, as in Latin *librum Petro dedit* (he/she gave the book to Petrus), where the oblique R argument *Petro* is in dative case, while the object *librum* is in accusative case.[5] The *secundative* strategy is not described in the UD guidelines, but it is natural to assume that the core R argument is annotated *obj* (since the *iobj* relation normally requires the presence of an *obj* argument in the same clause), while the oblique T argument is annotated *obl*.

Interim Summary

The upper part of Table 1 summarizes the UD treatment of basic voice constructions by showing how prototypical semantic roles are mapped to syntactic relations (with the two different treatments proposed for transitives with ergative alignment).

4 Non-Basic Voice Constructions

Non-basic voice constructions are clausal constructions used to express a non-prototypical combination of the topicality of referents and the participant roles those referents play in the event denoted by the predicate.

4.1 Passive–Inverse Constructions

A passive–inverse construction expresses a situation where the P referent has higher topicality than the A referent (Croft, 2022, p. 252). In the English passive construction in (8b), the P argument (*he*) is coded like the A argument (*she*) in the prototypical active construction in (8b), while the A argument (*by her*) is oblique. In the Algonquian inverse construction in (9b), from Wolfart and Carroll (1981), the P argument is again coded like the A argument in the direct construction in (9a), but the A argument is now coded as the P argument in the more prototypical construction.

(8) a. she took him to school
b. he was taken to school (by her)

(9) a. ni-wāpam-ā-wak
1-see-DIR-3PL
'I see them'
b. ni-wāpam-ikw-wak
1-see-INV-3PL
'they see me'

These are only two of the many strategies used in passive–inverse constructions in the world's languages. For further discussion, see (Croft, 2022, pp. 256–263).

4.2 Antipassive Constructions

Antipassive constructions involve a P argument with lower topicality than in a basic transitive clause. Such constructions are common in ergative languages, where the P argument is demoted to an oblique and the A argument takes over the absolutive encoding. Example (10), from Patz (2002), illustrates the antipassive construction in Kuku Yalanji.

(10) a. nyulu dingkar-angka minya-Ø nuka-ny
3SG.NOM man-ERG meat-ABS eat-PST
'the man ate meat'
b. nyulu dingkar-Ø minya-nga muka-ji-ny
3SG.NOM man-ABS meat-LOC eat-ANTP-PST
'the man had a good feed of meat'

[5] A dative case argument may be treated as a core argument, hence *iobj*, if other criteria point to it being core, notably if it is indexed on the verb.

Example (10) illustrates the oblique P strategy, which is also found in an English example like *the dog chewed the bone* versus *the dog chewed on the bone* (although without overt coding on the verb). Other common strategies in antipassive constructions are the omitted P strategy (*she ate a sandwich* versus *she ate*) and different types of noun incorporation (Croft, 2022, pp. 266–270).

4.3 Causative Constructions

Causative constructions add an external causer (C), universally encoded as a subject core argument. The encoding of the ordinary subject, the causee, depends on what strategy is used, and sometimes also on whether the base clause is transitive or intransitive. Many languages use a *complex predicate* strategy, as in the English examples in (11), where the causee becomes the direct object and everything else stays the same.

(11) a. she made him cry
b. she made him write the letter

Turkish instead uses a *simple predicate* strategy, with overt coding on the verb, as shown in (12), from Comrie (1989). In (12a), the base clause is intransitive and the causee is expressed as an object with accusative encoding; in (12b), the base clause is transitive and the causee is expressed as an oblique with a dative flag.

(12) a. Ali Hasan-ı öl-dür-dü
Ali Hasan-ACC die-CAU-PST
'Ali killed Hasan'

b. Dişçi mektub-u müdür-e imzala-t-tı
dentist letter-ACC director-DAT sign-CAU-PST
'the dentist made the director sign the letter'

4.4 Applicative Constructions

In applicative constructions, a peripheral participant is encoded as a core argument, usually as an object, and the object of the corresponding prototypical transitive clause may be encoded as an oblique. This is illustrated with a Hungarian example in (13), from Moravcsik (1978).

(13) a. János fák-at ültetett a kert-be
John trees-ACC planted the garden-into
'John planted trees in the garden'

b. János be-ültette a kerte-t fák-kal
John APPL-planted the garden-ACC trees-with
'John planted the garden with trees'

Hungarian uses a simple predicate strategy, with overt coding on the verb, but it is also common to use a complex predicate strategy for applicative constructions, in particular a serial verb strategy.

4.5 UD Annotation

Passive–Inverse Constructions

Passive constructions are annotated in UD by attaching the passive subject to the verb with the subtype relation *nsubj:pass* to indicate that it expresses the argument role associated with the direct object in the corresponding transitive clause. The agent phrase, if present, is annotated using the subtype *obl:agent*. If the verb is overtly marked for the passive voice, it carries the feature Voice=Pass; if the the passive is a periphrastic construction, the auxiliary may instead be annotated with the subtype *aux:pass*. Inverse constructions like the one in (9b) are not discussed explicitly in the UD guidelines, but it seems straightforward to use the subtypes *nsubj:pass* and *obj:agent* recommended for transitive clauses in (some) languages with ergative alignment, with the feature Voice=Inv on the verb.

Antipassives

Antipassives are not explicitly discussed in the UD guidelines, but the oblique P and omitted P strategies can be straightforwardly annotated using the existing guidelines.[6] The treatment of noun incorporation in UD is a more controversial issue, which we will sidestep in this paper. We refer the interested reader to Tyers and Mishchenkova (2020) for a discussion and a proposal.

Causatives

For causatives with the simple predicate strategy, UD recommends using the subtypes *obj:caus* and *iobj:caus* for the causee, as shown in Figure 3 for the Turkish examples in (12) (with the feature Voice=Cau on the verb). The use of the *iobj* relation here is unexpected, given that the argument has an oblique encoding, and the subtype *obl:caus* would seem more natural. For the complex predicate strategy, illustrated by the English example (11), the causee will normally be annotated with the *obj* role (without subtype), while the second verb will be assigned the *xcomp* relation.

[6] In the former case, the demoted P argument is assigned the *obl* relation; in the latter case, it is simply dropped.

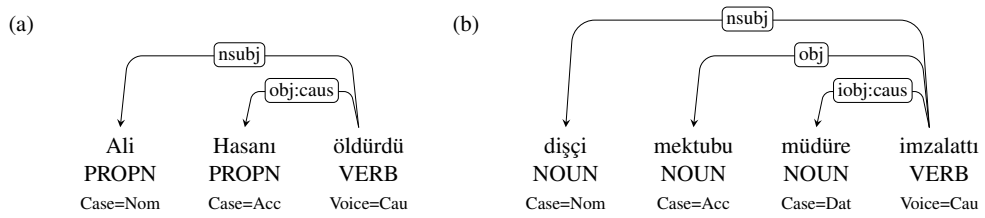


Figure 3: UD annotation of causative constructions.

Applicatives

Applicatives do not appear in the official UD guidelines, but there is a short discussion in de Marneffe et al. (2021) of ditransitive applicatives in Swahili, where it is recommended to use the *iobj* relation for the promoted argument if it is indexed by the verb. For the Hungarian example (13b) it seems natural to use *obj* for the promoted argument and *obl* for the demoted one. In addition, one could envisage a feature Voice=App on the verb, but no such feature currently exists in UD.

Interim Summary

The treatment of nonbasic voice constructions in UD is summarized in the lower part of Table 1. We have, however, only included constructions for which there are official guidelines.

5 Discussion

Our review has shown that the UD annotation framework can in principle represent all the major constructions and strategies for verbal predication discussed in Croft (2022), even though not all nonbasic voice constructions are treated explicitly in the current documentation of the UD guidelines. These guidelines are summarized in Table 1, which can be regarded as a blueprint for the UD constructicon of verbal predication constructions.

However, we have also observed a few cases where the UD treatment does not quite align with comparative concepts from typology, and sometimes arguably even conflicts with basic principles of UD itself. One such case is the treatment of transitives in ergative languages, where de Marneffe et al. (2021) advocates a mixed analysis, which is sometimes based on topicality, sometimes on morphosyntactic encoding, specifically case alignment. Another case is the analysis of ditransitive clauses with neutral alignment, where the use of the *iobj* relation appears to be motivated on different grounds than other core argument relations. Finally, we note that the use of subtypes to mark non-prototypical argument realizations can be improved with respect to systematicity and naming conventions. Nevertheless, we believe that, with relatively small adjustments, the guidelines can be made globally coherent and consistent with basic UD principles as well as findings from linguistic typology. We will now try to outline these modified guidelines and their motivation.

A cornerstone of UD is the assumption that the core-oblique distinction, albeit not completely unproblematic, is a better foundation for morphosyntactic annotation than the argument-adjunct distinction (de Marneffe et al., 2021, pp. 266–268). The basis for distinguishing core arguments in a given language is the encoding of the two arguments in a prototypical transitive clause; any argument that uses the same encoding as one of these is core; any argument that uses a different encoding is oblique.

The basis for assigning specific syntactic relations to core arguments in *basic* voice constructions is topicality, with the *nsubj* relation reserved for the most topical argument and the *obj* relation for the second most topical argument. It follows that the single S argument in intransitive clauses is *nsubj*.

In transitive clauses, we assume that the topicality hierarchy is A > P, which means that the A argument is *nsubj* and the P argument *obj*, regardless of case marking or other coding properties, and all other arguments are *obl*. This analysis naturally carries over to other event types like motion events: F = *nsubj*, G = *obj* in uncaused motion (*she entered the cave*); A = *nsubj*, F = *obj*, G = *obl* in caused motion (*she chased them into the cave*). For experiential events, the analysis mirrors the encoding in the prototypical transitives, which means that M = *nsubj* and X = *obj* in the causative construal (*she frightens them*) and vice versa in the attending construal (*they fear her*).

In ditransitive clauses, we assume that topicality reflects the force-dynamics (A > T > R), which means that A is *nsubj* and that T is *obj* if it is realized as a core argument; the expected realization of

R is *obl*, which makes the indirective alignment the basic voice construction for ditransitives. We will therefore treat the neutral and secundative alignments as nonbasic voice constructions (more precisely as applicative constructions), which obviates the need for the *iobj* relation.

In *nonbasic* voice constructions, which by definition involve some kind of mismatch between topicality and encoding, we use subtypes to indicate deviances from prototypical argument realizations. Here we propose a new subtyping system based on the argument roles used in linguistic typology, including at least *:s*, *:a*, *:p*, *:t*, *:r*, and *:c* (*c* for causer). We believe that this will be a more expressive and coherent subtyping system than the current use of *:pass*, *:agent*, and *:caus*, which mixes different naming conventions (constructions vs. roles) and where especially *:pass* has a misleading name as it covers more than just passives. Given these subtypes, we can annotate nonbasic voice constructions transparently as follows:

- **Passive–Inverse:** The P/T/R argument is *nsubj:p*/*nsubj:t*/*nsubj:r*, and the A argument is *obj:a* or *obl:a*, depending on strategy.
- **Antipassive:** The A argument is *nsubj* and the P argument is *obl:p*.
- **Causative:** The causer is *nsubj:c*. If the base clause is intransitive, the S argument is *obj:s*; if the base clause is transitive, then the A argument is *obj:a* or *obl:a* and the P argument *obj* or *obl:p*, depending on strategy.
- **Applicative:** The A argument is *nsubj*, the P/T argument is *obj* or *obl:p*/*obl:t*, depending on strategy, and the promoted argument is *obj* with a subtype reflecting its role. A special case of this is a ditransitive with neutral or secundative alignment, where the R argument is *obj:r* (instead of *iobj*) and the T argument is *obj* (neutral) or *obl:t* (secundative).

A possible alternative to using role-based subtypes is to use a simpler system with only two general subtypes, *:high* and *:low*, which indicate that an argument has, respectively, higher or lower topicality than expected. The P argument would then be *nsubj:low* in a passive–inverse construction and *obl:high* in an antipassive construction. However, this would be a much less expressive system, which would make some nonbasic voice constructions indistinguishable (for example, inverse constructions and intransitive causatives).

Finally, and regardless of whether future versions of UD will adopt our proposed revisions of the annotation guidelines, there will be a need for additional morphological features to capture nonbasic voice constructions coded on the verb itself. This includes at least a feature or feature value for applicative constructions.

6 Conclusion

In this paper, we have taken another step towards a constructicon for UD, in the sense of Nivre (2025), by reviewing the way UD annotates constructions and strategies for verbal predication, following the taxonomy of Croft (2022), extending the previous work on reference and modification (Nivre and Croft, 2025). An overview of the constructicon is shown in Table 1, where we outline which syntactic relations are used to annotate different argument phrases across constructions and strategies. To this should be added the annotation of verbal predicates using part-of-speech tags, features and the *aux* relation, and of argument encoding through morphological features and the *case* relation, as described in Section 3.5.

Based on our review of the existing guidelines and annotation practices, we have also proposed some modifications to the guidelines that should be considered for future versions of UD. This includes modified guidelines for transitive clauses in (some) languages with ergative alignment, and for ditransitive clauses generally, as well as a proposal for a new subtyping system, which will make the annotation of nonbasic voice constructions more transparent. As stated in Nivre and Croft (2025), these proposals need to be evaluated also from other perspectives, since UD is designed as "a very subtle compromise between a number of competing criteria" (de Marneffe et al., 2021, p. 302), and the discussion also needs to be informed by a more comprehensive review of the UD framework, covering all major types of constructions and strategies. It is our goal to continue this review in a series of future publications.

Acknowledgments

This work received support from the CA21167 COST action UniDive and from the Swedish Research Council (grant no. 2022-02909). We thank André Coneglian, Leonie Weissweiler and three anonymous reviewers for comments and suggestions that helped improve the article.

References

John Beavers. 2011. On affectedness. *Natural Language and Linguistic Theory*, 29:335–370.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology*. University of Chicago Press.

William Croft. 2010. *Verbs: Aspect and Causal Structure*. Oxford University Press.

William Croft. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, 20(2):377–393.

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.

William Croft. in press. Comparative concepts. In Hilary Nesi and Petar Milin, editors, *Encyclopedia of Language and Linguistics*. Elsevier.

Jon P. Dayley. 1989. *Tzutujil Reference Grammar*. University of California Press.

Roger M. W. Dixon. 1994. *Ergativity*. Cambridge University Press.

Joseph H. Greenberg. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Grammar*, pages 73–113. MIT Press.

John Haiman. 2011. *Cambodian (Khmer)*. John Benjamins.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86:663–687.

Martin Haspelmath. 2011. On S, A, P, T, and R as comparative concepts for alignment typology. *Lingustic Typology*, 15:535–567.

Martin Haspelmath. 2015. Transitivity prominence. In Andrej L. Malchukov and Bernard Comrie, editors, *Valency Classes in the World's Languages*, pages 131–147. Mouton de Gruyter.

Arthur Lorenzi, Peter Ljunglöf, Ben Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig, and Ely A. Matos. 2024. MoCCA: A model of comparative concepts for aligning constructicons. In *Proceedings of the 20th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, pages 93–98.

Andrej L. Malchukov. 2005. Case pattern splits, verb types and construction competition. In Mengistu Amberber and Helen de Hoop, editors, *Competition and Variation in Natural Languages: The Case for Case*, pages 73–117. Elsevier.

Marie de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47:255–308.

Edith A. Moravcsik. 1978. On the distribution of ergative and accusative patterns. *Lingua*, 45:233–279.

Joakim Nivre. 2025. Constructions and strategies in Universal Dependencies. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 419–423.

Joakim Nivre and William Croft. 2025. Reference and modification in Universal Dependencies. In *Proceedings of the 8th Workshop on Universal Dependencies (UDW))*, pages 1–10.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 4034–4043.

Elizabeth Patz. 2002. *A Grammar of the Kuku Yalanji Language of North Queensland*. Australian National University.

Leonard Talmy. 1988. Force dynamics in language and cognition. *Cognitive Science*, 12:49–100.

Takasu Tsunoda. 1981. Split case-marking patterns in verb-types and tense/aspect/mood. *Linguistics*, 19:389–438.

Takasu Tsunoda. 1985. Remarks on transitivity. *Journal of Linguistics*, 21:385–396.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204.

David John Weber. 1989. *A Grammar of Huallaga (Huánaco) Quechua*. University of California Press.

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932.

C. J. Williams. 1980. *A Grammar of Yuwaalaraay*. Australian National University.

Christoph H. Wolfart and Janet F. Carroll. 1981. *Meet Cree; A Guide to the Cree Language*. University of Nebraska Press.

Linguistic Generalizations are not Rules: Impacts on Evaluation of LMs

Leonie Weissweiler **Kyle Mahowald**
The University of Texas at Austin
{weissweiler,kyle}@utexas.edu

Adele E. Goldberg
Princeton University
adele@princeton.edu

Abstract

Linguistic evaluations of how well LMs generalize to produce or understand language often implicitly take for granted that natural languages are generated by symbolic rules. According to this perspective, grammaticality is determined by whether sentences obey such rules. Interpretation is compositionally generated by syntactic rules operating on meaningful words. Semantic parsing maps sentences into formal logic. Failures of LMs to obey strict rules are presumed to reveal that LMs do not produce or understand language like humans. Here we suggest that LMs' failures to obey symbolic rules may be a feature rather than a bug, because natural languages are not based on neatly separable, compositional rules. Rather, new utterances are produced and understood by a combination of flexible, interrelated, and context-dependent *constructions*. Considering gradient factors such as frequencies, context, and function will help us reimagine new benchmarks and analyses to probe whether and how LMs capture the rich, flexible generalizations that comprise natural languages.

1 Introduction

How well do large Language Models (LMs) generalize beyond their training data? Much work on this question has presumed that generalizations require symbolic rules for syntax and semantics that generate acceptable new forms and compositional meanings. Rules are invoked to explain that if you learn a new modifier ('blonky') and a new count noun ('gravimin'), a compositional rule could predict that 'a blonky gravimin' is a gravimin that is blonky. In what follows, we use "rule" to refer to context-free generalizations that contain variables, to be instantiated by any instance of a general type, uninfluenced by frequency, similarity, or context (Pinker, 1999). Our focus here is on the use of a strict algebraic conception of rules, which we argue, underlies certain approaches to NLP evaluation, even though the notion of a rule is used variably in linguistics today, with several frameworks incorporating functional and/or frequency-based attributes into representations (e.g., Bresnan et al., 2007; Brehm et al., 2022; O'Donnell, 2015),

Because early statistical models (e.g., *n*-gram or Markov models) seemed unable to generalize fully or capture non-local dependencies (Chomsky, 1957), early on, rules seemed to many to be the only game in town for human language. After all, if a standard bigram model hadn't seen 'blonky gravimin' before, it would be unable to form a representation of it. Influential thinkers argued that neural networks, which did not involve rules, would never be appropriate models of human cognition for this reason (Fodor and Pylyshyn, 1988; Pinker and Prince, 1988; Marcus, 1998; Fodor and Lepore, 2002; Marcus, 2001; Calvo and Symons, 2014).

However, current LMs arose from statistical, distributional parallel models (Mikolov et al., 2013; Rumelhart et al., 1986) rather than rule-based natural language technologies. They do not rely on hard-coded rules, yet their ability to produce coherent, naturalistic language and respond appropriately is unparalleled by purely symbolic systems (Piantadosi, 2024; Goldberg, 2024; Weissweiler et al., 2023; Hofmann et al., 2025). GPT-4o, for example, not only recognizes 'a blonky gravimin' as a noun phrase, it explicitly offers several naturalistic interpretations, e.g., 'A person or act that awkwardly and absurdly pretends to be serious.'

Nonetheless, an assumption that generalizations are equivalent to rules continues to motivate many evaluations of syntax, meaning, and their compositional combination: e.g., Natural Language Inference (Bowman et al., 2015), Semantic Parsing (Palmer et al., 2005; Reddy et al., 2017), tests of binary grammatical acceptability (Warstadt et al., 2019; Dentella et al., 2023) and rule-based compositionality (Kim and Linzen, 2020). Together,

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 61–74
September 24, 2025,

such tasks made up more than half of the GLUE benchmark (Wang et al., 2018), created to evaluate language models on their skill at being "general, flexible, and robust." Lackluster performance on rule-based tasks in the early days of LMs was taken to imply that the models did not use language the way people do and were instead merely imitating shallow surface patterns (Bender and Koller, 2020; Kim and Linzen, 2020; Weißenhorn et al., 2022; Bolhuis et al., 2023). In a survey of 79 NLP researchers, McCurdy et al. (2024) reported that 87% believed LMs were not sufficiently compositional and a sizable proportion (39%) believed explicit discrete symbolic rules were required.

Evaluations of LMs' early challenges with algebraic or logical rules did expose certain shortcomings in their ability to reason abstractly and solve math problems (see e.g., Mahowald et al., 2024). At the same time, LM's concurrent ability to produce and respond to natural languages *naturalistically* is hard to overstate (e.g., Coil and Shwartz, 2023).

Mastering a natural language requires mastering a network of hundreds of thousands of context-dependent, gradient, flexible schemata (*constructions*, see §5), which often contain 'slots' that constrain their fillers and how those fillers are interpreted. Constrained slots allow for new combinations, flexibly adapted in context. For instance, the phrase '<time period> ago' can coerce a temporal interpretation of filler phrases that do not designate time periods (e.g., 'three rest stops ago'). Rather than rule-based compositionality, composition-by-construction allows constructions to contribute meaningfully to interpretation in ways that range from abstract to quite narrow and specific. Therefore, for LMs to use language like humans, they require interpretations that are far richer than rules can provide for thousands of collocations, conventional metaphors, idioms, and context-dependent interpretations. Even abstract grammatical patterns also regularly convey semantic and/or pragmatic information that restrict their contexts of use and interpretations. Since different languages and dialects provide speakers with different networks of constructions (ConstructionNets), cross-linguistic differences can be captured naturally.

We suggest that rule-based evaluations have been over-emphasized in the domain of natural language production and comprehension. Our goal is to emphasize the importance of recognizing context, frequencies, meaning and other gradient functional factors in modern evaluations of natural language.

We do not argue that no categorical rule exists in any language. If a categorical rule is needed, it can be treated as the limiting case, a fully abstract construction (Jackendoff, 2002). For instance, Jackendoff (2002) proposes a symbolic Verb + Particle rule for the syntax of English complex verbs. At the same time, the meanings of individual verb plus particle combinations are far from compositional by any general rule (e.g., one can *look up* a number or *look down on someone* but not *?look up on someone* nor *?look down a number*). Here we advocate for an increased focus on the extent to which and *how* LMs manage to produce and comprehend human-like natural languages in all their context-specificity and complexity.

Many of our theoretical points are not new, particularly in the domain of morphology. Neural network researchers have continuously argued in favor of a single representational system and against the usefulness of rules in the domain of words and inflectional morphology (e.g., Rumelhart et al., 1986; Rogers and McClelland, 2004; Elman, 2009; Christiansen and Chater, 1999; MacDonald et al., 1994; McClelland, 2015). While early work in Artificial Intelligence relied on algebraic rules (Minsky and Papert, 1969; Lenat, 1995), many researchers soon realized that rules were too brittle to scale up beyond highly restricted domains such as artificial block worlds (Winograd, 1980).

Our contribution is to review leading paradigms used in LM evaluations of syntax (§2), semantics (§3), and compositionality (§4). We argue that, while these paradigms have been fruitful, they inherit from a tradition that was overly focused on rules, hierarchy, compositionality, and a binary notion of grammaticality. We briefly characterize how these assumptions arose and how they are baked into evaluations. We argue that **evaluations that presume categorical and strictly compositional language ignore some of the richest elements of human language**. We review construction-based and gradient functionalist approaches to language, arguing that this tradition points to certain lacunae in existing evaluations and open up new possibilities for evaluating natural language understanding. Early work in this direction has already included more nuanced metrics, measuring gradient judgments and context-dependent interpretations (e.g., Juzek, 2024; Hu et al., 2024).

2 Syntactic Rules in LM Evaluations

Evaluations of the syntactic capabilities of LMs have frequently assumed a binary categorical notion of grammaticality, which is then used to create datasets for evaluation. Below, we discuss several such cases, attempting to make these assumptions explicit to show their limitations.

Grammaticality Judgment Tasks Human judgments on sentences are gradient rather than binary, and demonstrably depend on frequency, plausibility, complexity, memory demands, potential alternatives, and context (Grodner and Gibson, 2005; Schütze and Sprouse, 2013; Robenalt and Goldberg, 2015; Gibson and Hickok, 1993; Fang et al., 2023). The amount of exposure to written language or linguistic theory also influences people's judgments. For instance, Dąbrowska (2010) found that laypeople's judgments on sentences containing long-distance dependencies were more sensitive to lexical content than linguists' judgments were. Even sentences included in linguistic textbooks, which one might presume to have clear-cut judgments, in reality are judged gradiently by people (Juzek, 2024). Nonetheless, LMs' language skills are often evaluated on binary grammaticality judgments on sentences (Dentella et al., 2024, 2023; Warstadt et al., 2019).

The fact that human judgments are gradient can have profound consequences on evaluations. For instance, Dentella et al. (2023) compared humans and LMs against predetermined binary acceptability labels, reporting that LMs' performance correlated poorly. However, comparing gradient perplexity measures with the same human judgments revealed a strong positive correlation (Hu et al., 2024). Using perplexity measures for models (as well as allowing humans to provide ordinal or gradient judgments) is a step in the right direction (Hu et al., 2024; Juzek, 2024).

Dependency Parsing as Evaluation Parsing text for universal dependencies (UD, de Marneffe et al., 2021) has become a well-established task for evaluating models (Zeman et al., 2017, 2018), and since Hewitt and Manning (2019) showed BERT (Devlin et al., 2019) to be somewhat skilled in UD, it has became the default operationalization of syntax in the NLP world (Amini et al., 2023; Kryvosheieva and Levy, 2025; Müller-Eberstein et al., 2022) and in discussions of inductive biases (Lindemann et al., 2024; Glavaš and Vulić, 2021). UD annotations are partially determined by semantics and they are based on lexical items, which makes them closer to the approach advocated here rather than abstract phrase structure rules. However, UD analyses presume a universal set of grammatical relations, which is problematic, because not all languages employ the same constructs. That is, there is no universally valid way to define or test for the syntax of nouns, verbs, adjectives, subjects, or direct objects (e.g., Croft, 2001). Moreover, UD requires an asymmetric relationship between a 'head' and dependent, yet the long tail of language includes headless constructions (e.g., *the Xer, the Yer* construction) (Michaelis, 2003) and co-headed constructions (e.g., phrasal verbs, conjunctions, idioms). Therefore, UD annotations need to be determined for individual languages and need to allow for non-headed or co-headed cases to align well with formal aspects of natural languages.

3 Semantic Rules in LM Evaluation

Formal logic was developed as a branch of mathematics, used to prove mathematical and philosophical theorems and identify provability gaps (Frege, 1918; Russell, 1905; Gödel, 1931). It was based on algebraic rules operating on categorical and broadly defined categories. Notably, many logicians did not generally assume nor endorse using formal logic to represent the meanings of natural language utterances (Carnap, 1937; Baker and Hacker, 1986), recognizing that natural languages differ from formal logic in many ways.[1] For instance, logic treats *and* and *but* as equivalent. It does not provide a natural way to capture commands or questions (Austin, 1975), nor does it naturally distinguish presuppositions from assertions (Strawson, 1967). Finally, formal logic is not intended to capture effects of context (Wittgenstein, 1953; Russin et al., 2024).

Yet the assumption that natural language semantics can be modeled by formal logic has been made in the design of certain classic LM understanding benchmarks. Below, we review some instances and discuss their connection to formal semantics.

Natural Language Inference Natural Language Inference tasks label the second of two sentences as an entailment, contradiction, or neutral, and this NLI task was originally used to train models (Wang et al., 2019; Dagan et al., 2006; Nie et al., 2020).

[1] Some logicians did advocate for using formal logic for natural language (Tarski, 1944; Montague, 1970, 1973).

Today, NLI is used as a zero-shot evaluation metric to assess natural language understanding (Zhou et al., 2024; McCoy et al., 2019). In introducing the Stanford NLI corpus, Bowman et al. (2015) state, "The semantic concepts of entailment and contradiction are central to *all* aspects of natural language meaning. ") (emphasis added, see also Katz, 1972; van Benthem, 2008). While in the same paper, Bowman et al. (2015) acknowledge that judgments depend on many factors, such as commonsense knowledge, this fact is generally overlooked in papers that use NLI as a task to evaluate LMs' general understanding.

Necessary and plausible inferences *are* a critical aspect of natural language understanding. However, they are highly dependent on the interlocutors' general communicative goals. We aim to make sense of others' messages, so we assume others are trying to be relevant and helpful and do our best to assign coherent meanings to all utterances (Grice, 1975). For example, outside of logic classes or heated arguments, people rarely conclude that two statements made by the same person are contradictory. If someone utters: 'The boy is depressed but he is not DEPRESSED', listeners do not throw up their hands and shout 'contradiction!'. Instead, they may infer that the boy in question is only somewhat, and not extremely, depressed, or ask to learn more. People also assign interpretations to statements in ways that differ from what formal logic would predict (e.g., 'run fast and you've got this' or 'If it snows, it snows.' NLI tasks that rely on judging contradictions or entailments may over- or under-estimate how well LMs' understanding of natural language aligns with humans', particularly when binary judgments are required (cf. Dentella et al., 2024).

Evaluation metrics need to take humans' communicative goals into account and allow for gradient and context-dependent interpretations. An example of the type of evaluation we endorse can be found in the underappreciated NOPE testbed. Parrish et al. (2021) selected 10 distinct constructions that trigger presuppositions and curated 100+ instances of each one, based on naturally occurring examples. Each stimulus includes two preceding sentences for context. The authors then collected gradient judgments from human raters, allowing them to use their 'background knowledge about how the world works' and compared the accuracy of several models, with several controls in place. This strikes us as a highly valuable blueprint for modern evaluations of LMs.

Semantic Parsing Banarescu et al. (2013) introduced abstract graphical meaning representations (AMR) for sentential meaning that importantly includes aspects of lexical semantics. It was created to offer a repository of structured meanings to be used for evaluating understanding in LMs (Li et al., 2023; Qiu et al., 2022; Shaw et al., 2021, see also §4). Yet work on AMR concedes that it ignores so-called 'syntactic idiosyncracies.' For example, 'he described her as a genius' and 'his description of her: genius' are assigned the same AMR. Yet the former is unambiguously about her intellect, while the latter may instead be used to compliment *his cleverly tactful description*. More generally, simplifications of distinctions made in a natural language can be expected to result in lost meaning, since two utterances are rarely interchangeable in all contexts. Focusing on subtle but important differences in meaning offers an opportunity to design more challenging linguistic evaluations of LMs.

4 Compositionality

As computer coding languages became more and more widespread, rule-based syntax and semantics took root in linguistics. A Principle of Compositionality states that the semantics of a sentence is determined by the meanings of the words and the syntactic rules used to combine them (Montague, 1970; Partee, 1984; Dowty, 1979; Jackendoff, 1992; Fodor and Lepore, 2002). It is intended to be a bottom-up process: syntactic rules combine words, which have determinant meanings. Fodor and Pylyshyn (1988) make this clear: "a lexical item must make approximately the same semantic contribution to each expression in which it occurs". That is, context may not influence the interpretation of words in a top-down manner; therefore downstream inferences are required to address the fact that interpretations *do* depend on context. Realizing this, like Carnap and Frege before him, Fodor and Pylyshyn (1988) acknowledge: "It's uncertain exactly how compositional natural languages actually are."

Nonetheless, compositionality is often taken as a truism, based on the standard argument for it summarized below.

> People tend to agree on the interpretation of new sentences. ⇒ There must be some set of rules that determine the meaning of new sentences.

Note that one can agree with the premise without accepting the consequent. In particular, people generally agree on the interpretations of pointing gestures and novel words as well as sentences, and yet shared interpretations *must* be gleaned from non-linguistic context in the case of pointing gestures, and from a combination of linguistic and non-linguistic context in the case of novel words. Shared interpretation of sentences likewise comes in part from linguistic and non-linguistic context.

Consider the sentence, 'the Persian cat is on the mat.' If the speaker's goal is simply to help someone find the furball, there need be no commitment to the cat being a thoroughbred Persian breed nor to the cat being wholly on, rather than adjacent to, the mat. Or, comprehenders may appreciate the statement is ironic if the cat is hairless.

Cases that might seem amenable to a rule often turn out to require a good deal of item-specific memory. For instance, a compositional rule involving set-intersection may seem appealing for '<color term> noun' combinations in the domain of artificial block worlds (e.g., a green cube is something that is both a cube and green). However, violations of such rules abound: green tea is more yellow than green, and Cambridge blue is actually green. Even more common are instances that evoke richer meanings than predicted by any algebraic rule: e.g., a green light implies that forward motion or progress is permitted, and a green card provides a path toward citizenship in the US (or ought to). The meanings of familiar collocations are typically not fully determined by general compositional rules, and novel cases can be interpreted on analogy to familiar cases rather than according to some very general rule. For instance, if "flam" is interpreted to mean any kind of event or action, *a green flam* is likely to be interpreted to imply an eco-friendly or beginning-level event. Representing only the rule-compliant cases in evaluations can therefore lead to the wrong conclusions. A more comprehensive evaluation paradigm should take into account how people actually interpret familiar and novel cases.

Another issue is that rules massively overgenenerate. That is, rules predict all manner of odd locutions (Pawley and Syder, 1983; Sag et al., 2002): e.g., 'Meeting you is pleasing to me'; 'The tall winds hit the afraid boy'; 'Explain them the problem.' Humans are sensitive to the frequencies of various types of word combinations and judge formulations unnatural if there exists a more conventional way to express the intended message in context (e.g., Goldberg, 2019).

Evaluating LMs for Compositionality Compositionality benchmarks combine elements from syntactic and semantic evaluations. Kim and Linzen (2020)'s compositional generalization challenge (COGS) tested whether models could translate any sentence generated by a small set of syntactic rules into formal semantics. For instance, trained on representations of 'the girl,' 'the cat,' 'the hedgehog,' 'the cat loves the girl,' 'the hedgehog sees the cat,' and so on, the model was tested on how well it predicted a formal semantic representation of 'The girl loves the hedgehog.' However, note that if 'mosquitoes' is substituted for 'the cat,' different interpretations of 'love' are evoked ('Mosquitoes love the girl' vs. 'The girl loves mosquitoes'), not to mention different degrees of plausibility. The authors also anticipated generalizations from sentences like 'Jane gave the cake to John' to 'Jane gave John the cake,' and the models were found to perform poorly. Yet the two sentences differ in terms of information structure (Bresnan and Ford, 2010) and the relative frequencies and similarities of verbs witnessed in each version (Leong and Linzen, 2024; Ambridge et al., 2014; Hawkins et al., 2020). Thus, while an evaluation of this kind can capture something about how humans interpret automatically generated sentences in an experimental context, focusing on this type of task may distort our view of how well LMs handle natural language in the wild.

Other compositionality benchmarks adopt NLI tasks, which commonly presume interpretation is determined by rules. For example, in the context of robotic agents interpreting instructions, Lake and Baroni (2018, p.1) state:

> Humans can understand and produce new utterances effortlessly, thanks to their compositional skills. Once a person learns the meaning of a new verb 'dax', he or she can immediately understand the meaning of 'dax twice'...

The robotic agents struggled to interpret the rule-based command, though it was appropriate in the narrow domain tested. Notably, the rule does not apply to natural language generally. For instance, unbounded actions are not countable, so if 'twice' appears at all, it is likely followed by a comparative phrase (e.g., 'work twice as hard'), which has a very different meaning than performing an ac-

tion two times. Other cases require knowledge of specific combinations: 'to think twice' means 'to hesitate' and 'going twice' tends to evoke the context of an auction. Familiar phrases with meanings not fully captured by compositional rules are common: By one estimate, we learn tens of thousands of them (Jackendoff, 2002). Importantly, we tend to agree on their interpretations, even though each means something more or different than predicted simply by the words and their syntactic combination. In this way, phrasal combinations regularly involve subregularities or item-specific interpretations not predicted by general algebraic rules.

Another example comes from the seemingly innocuous algebraic rule: "If X is more Y than Z, then Z is less Y than Z, irrespective of the specific meanings of X, Y, and Z" (Dasgupta et al., 2020, :5). This is meant to capture that 'Anne is more cheerful than Bob' should both contradict 'Anne is less cheerful than Bob', and entail 'Bob is less cheerful than Anne.' NLI models that failed to draw these inferences were considered lacking. Yet natural language rarely relies on free variables. The content of X, Y, and Z matters. No one would infer that because Anne_x is more cheerful_y than careful_z, that 'Careful_z is less cheerful_y than Anne_x.' Perhaps more importantly, if a speaker uttered 'Anne is higher than Bob and Bob is higher than Anne,' listeners would likely infer either that Bob climbed above Anne in the time it took to utter the first clause or that Bob has been smoking. We have so far argued that an overemphasis on symbolic abstract rules for natural languages can lead to evaluations of natural language that are not aligned with humans. Below we suggest an alternative approach to language, which we argue helps refocus evaluations on interesting new research questions.

5 The Constructionist Approach

This section briefly explains the constructionist approach to language, which conceives of a language as a vast network of interrelated *constructions*, of varying size and complexity. This differs from a perspective that treats languages as a set of sentences generated by a small set of algebraic rules. We suggest a change of perspective about the nature of language, not a mere substitution of the units on which some type of rules operate. That is, certain traditional evaluations were far too limited in requiring models to adhere to strict compositionality, when humans do not. At the same time, the constructionist approach encourages stringent evaluations by testing whether models capture the gradient and function-sensitive patterns that characterize natural languages.[2] The approach encourages us to broaden our view of language and linguistic evaluations of LMs.

(Partially-filled) Words, Common & Rare Schemata as the same type of Representations A 'construction' is any learned association between a formal pattern and a range of related functions. This simple definition treats words, idioms, rare *and* common grammatical patterns as constructions. As a result, the lexicon and syntax are not treated as distinct or modular systems. This allows the many parallels between them to be easily captured. It also allows a natural way to allow for the diversity found in the world's languages, in which more or less information is encoded in a single word. Formal attributes of constructions include phonology, grammatical categories, word order, discontinuous elements, specific words or morphemes, and/or intonation. Any construction may include one or more constrained open 'slots'.

A Wide Range of Functions Considered Jointly with the Forms Constructions' functions vary widely: words, collocations, and idioms convey rich, specific, contentful meaning. A plethora of other constructions are productive but constrained in a variety of semi-specific ways; argument structure constructions convey 'who did what to whom'; discourse structuring constructions indicate which parts of an utterance are at-issue or backgrounded. A range of constructions exist to ask questions, express surprise or disapproval, greetings or gossip. Construction can be associated with specific registers, genres, and/or dialects. The constructionist commitment to considering semantics jointly with syntax represents a more comprehensive understanding of their interactions, which can help develop tests that evaluate both.

Sensitivity to Similarity and Frequency Language users are sensitive to the frequencies of constructions. For instance, the passive construction is used far more frequently in Turkish than English and young Turkish speakers learn the construction far earlier than English-speaking children (Slobin, 1986). Constructions are also influenced by simi-

[2]For more comprehensive introductions to the constructionist approach, or 'Construction Grammar', see Hoffmann and Trousdale (2013) and Hoffmann (2022).

larity: Instances of a construction prime instances of the same or closely related construction (e.g., Du Bois, 2014; Pickering and Ferreira, 2008). Constructionist approaches take this to be a core aspect of language and language learning, rather than an inconvenience or afterthought. This leads to a de-emphasis of definitional boundaries and an organic incorporation of fuzzy boundaries and prototypicality effects.

Productive Constructions May Include Fixed Lexical Units Syntax, semantics, and morphology are interrelated rather than assigned distinct levels. This is useful because even productive hierarchical constructions often include particular words and semantic constraints. For example, an English construction that implies real or metaphorical motion allows a wide range of verbs but requires the particular noun 'way' ('He charmed his way into the meeting.').

Interrelated Network, Not an Unstructured Set Unlike rules, which are commonly presented as unstructured lists, constructions comprise a network of interrelated patterns. This allows for the fact that each language includes families of related constructions. It also allows for the simple fact that productive constructions simultaneously co-exist with specific conventional instances. For instance, the English 'double object' construction is productive, and speakers are also familiar with dozens of conventional instances (e.g., 'give <someone> the time of day', 'throw <someone> a bone').

More Maximal than Minimal A ConstructionNet includes words as well as grammatical patterns, and lossy instances are included as well as generalizations across instances, as just mentioned, which provides some redundancy. There is no reason to restrict the complexity of constructions or their descriptions more than is warranted by psychological and linguistic evidence.

Construction Slots Are Constrained The open 'slots' of constructions are constrained in a wide variety of ways. For instance, the English double-object construction can appear with a wide range of verbs, but prefers simple verbs to those that sound Latinate (e.g., 'She told them something' vs. 'She proclaimed them something'). The English comparative suffix '-er' (e.g., 'calmer', 'quicker') is available for most single-syllable adjectives that allow a gradient interpretation, but it is not used with past participle adjectives (? 'benter').

An Example Consider 'X is the new Y'. It is productive and can be used to create new utterances, e.g., 'Semiconductor chips are the new oil.' As is typical of productive constructions, the generalization co-exists with several familiar instances (e.g., '50 is the new 40'; 'Orange is the new black'). The construction is not an algebraic rule: Its slots, indicated by X and Y, are not variables that range freely over fixed syntactic categories. Instead, 'X' must be construed (playfully) as currently functioning in the culture as 'Y' used to. Therefore, not all combinations of slot fillers make sense: (e.g., ? 'Orange is the new oil'). Adding a parallelism constraint between X and Y is insufficient since '103 is the new 101' would also require an unusual context to make sense. Finally, instances of the construction are not amenable to a general compositional rule, nor can they be translated into formal logic. Either approach would presumably treat 'Orange is the new black' as equivalent to 'Black is the old orange,' which does not conventionally evoke the same meaning.

6 Implications Beyond Natural Language

Outside of natural language, even in domains that are rule-like by design, rule-based interpretations are sometimes lacking, potentially due to the fact that natural language is used by people when discussing these domains. For instance, LMs have been found unreliable at drawing the following inference, which the authors dubbed the *reversal curse*: "if 'A is B' [...] is true, then 'B is A' follows by the symmetry property of the identity relation" (Berglund et al., 2023, p. 2).

Why are LMs prone to the reversal curse? Although the quote above is stated in natural language, it does not apply to natural language sentences, which are rarely reversible without a different interpretation: e.g., 'A mental illness is the same as a physical illness' means something very different than 'A physical illness is the same as a mental illness' (see also Tversky, 1977; Talmy, 1975). Even simple conjunctions are not generally reversible in natural language. For instance, 'night & day' and 'day & night' are both acceptable, but their interpretations differ: the former conveys a stark contrast (e.g., 'as different as night and day'), the latter suggests a relentless activity or process (e.g., 'he worried day and night'). In summary, it is perhaps reasonable to expect truly symmetric knowledge to be reversible. But LMs are trained

on natural language, which is not symmetric.

7 New Directions for Evaluation

Natural languages involve complex and context-sensitive systems of constructions, which vary from being wholly fixed to highly abstract and productive. Constructions are combined when a unit, potentially itself composed of constructions, fills a slot in another construction. Viewing language as a system of constructions rather than words and rules may fundamentally change how the successes and failures of models are construed, and new goals and questions come into focus. The complexity of constructions with respect to gradience in frequencies, functions, slot constraints, and prototypicality can be used to develop evaluations that demand the same complexity from LMs found in natural languages.

A caveat is required for low-resource languages, where rule-based linguistic evaluations (e.g., Jumelet et al., 2025) can be useful. More generally, evaluations should meet models where they are: if the representational complexity of an LM is restricted, restricted types of evaluations are required. But when evaluating LMs on high-resource languages, richer evaluations are appropriate. Specifically, we recommend the following.

When possible, use a variety of naturalistic sentences rather than sentences generated by a template that presupposes grammatical rules with interchangeable vocabulary items, as is done, e.g., by Multi-NLI (Williams et al., 2018). The idea that sentences can be constructed by subbing random lexical items into templates often misses lexical subtleties that are an important part of natural language. Instead, ecologically valid stimuli can be collected or adapted from natural corpora and normed for naturalness and plausibility. Since human judgments are highly context-dependent, benchmark tasks should also vary contexts systematically (see, e.g., Ross et al., 2024; Parrish et al., 2021).

Collect human assessments that allow for gradient context-sensitive interpretations that appeal to learned constructions. Evaluating LM competence on individual constructions requires assessing both acceptability judgments and interpretations from humans to draw appropriate comparisons.

We also need to be sensitive to the implications people and LMs draw from instructions and the testing context. **Do not give instructions to human-evaluators (or models) in ways that make the results a foregone conclusion.** If people are instructed to interpret 'red X' as 'X that is red for any X,' they are capable of doing so; this may reflect the instructions, not their natural intuitions. In natural contexts, people understand that red grapefruits are closer to pink, red hair is more orange, a red book may be about communism, and crossing a red line may have consequences.

The items included in testing also influence interpretations by people and models, by providing context. For instance, if certain pairs of items are mismatched (e.g., "The cup is green." "The cup is blue.") while others are matched, people can infer which ones are intended to be contradictory. LMs, like people, are now capable of generalizing by rule when tasked to do so. For instance, Lampinen et al. (2025) found Gemini 1.5 Flash (Team et al., 2024) avoided the reversal curse, achieving 100% accuracy, when the Berglund et al. (2023) dataset was provided to the LM as context.

A variety of items, participants, and contexts ought to be valued as much as a variety of models. It has long been recognized that real words, phrases and sentences vary in an open-ended number of ways (Clark, 1973). So care must be taken to include a variety of stimuli items. Because linguistic meaning is deeply tied to local context, even seemingly similar sentences can have very different interpretations in ways that depend on context. Different subgroups of participants may perform differently so distinct dialects should be taken into account.

The most interesting questions may no longer be *whether* LMs are skilled at producing and responding to natural languages, but *how* they achieve such remarkable skills. As is familiar from the lexicon, constructions comprise an interrelated network. We can now **how relationships between constructions are picked up by LMs.** For instance, Misra and Mahowald (2024) have demonstrated that even when all instances of a rare non-compositional construction are ablated from training data, non-trivial learning of the construction remains, enabled by the presence of related constructions in training. Moreover, nearly every productive construction coexists with at least a few formulaic instances, and LMs offer ways to test various theoretical perspectives on the nature of those relationships. These and other newer ways of probing LMs are possible,

and our toolkit will only grow.

As discussed by Weissweiler et al. (2023), investigating whether LMs distinguish subtle meaningful differences between constructions is another important direction. Recent work on this has included Weissweiler et al. (2022), who found LMs reliably discriminated instances of the English Comparative Correlative from superficially similar expressions. Tayyar Madabushi et al. (2020) tested a dataset of automatically induced constructions and reported that BERT (Devlin et al., 2019) could determine whether two sentences contained instances of the same construction. As mentioned earlier, Tseng et al. (2022) showed that LMs gradiently predict appropriate slot fillers. Li et al. (2022) probed RoBERTa's implicit semantic representations of four argument structure constructions (ASCs) and found similarities in behavior in the model and a sorting task done by humans. However, Zhou et al. (2024) found LMs failed to distinguish entailment differences between the causal excess construction (e.g., 'so heavy that it fell') and two structurally similar constructions ('so happy that she won'; 'so certain that it rained').

8 Conclusion

Generalization is a key component of human language—and a big part of why LMs are successful at processing language. But we have argued that evaluations of the linguistic abilities of LMs are too often based on an assumption that generalization requires algebraic rules operating on words. Natural languages are not Lego sets. Instead, language involves flexible combinations of rich and varied constructions of differing sizes, complexities, and degrees of abstraction, which differ from algebraic rules in many ways. By designing new evaluations that accurately reflect the complexities of language, we can avoid under- or overestimating language models. The extent to which LMs produce and interpret combinations of constructions has only begun to be explored. We believe future progress lies not in asking whether LMs obey abstract rules, but in probing what kinds of constructions they learn, how they relate them, and how those structures guide novel interpretation and production. In doing so, we may better capture what it truly means to comprehend and use language.

Limitations

While we have aimed to discuss benchmarks and evaluations in ways that reflect the historical trajectory as well as the present-day landscape, evaluations of LMs are continually developing. We feel the dominant paradigms have and continue to be based on data generated by rules and evaluated without regard for context effects, gradience, or semantic nuance, but we are keenly aware that we have likely overlooked metrics that go beyond rule-based evaluations (e.g., Parrish et al., 2021).

We recognize the growing work in multilingual evaluations, which are inherently valuable (Mueller et al., 2020; Jumelet et al., 2025; Kryvosheieva and Levy, 2025). The current perspective applies to all natural languages, but comparative work is not the focus of the current perspective, and we use English examples for the sake of easy comprehension and brevity.

Acknowledgments

We thank Najoung Kim, Kanishka Misra, and Will Merrill for helpful discussions and feedback. We are grateful to audiences at the NSF-sponsored New Horizons in Language Science workshop and the Analytical approaches to understanding neural networks summmer school sponsored by Simon's Foundation for helpful feedback. Leonie Weissweiler was supported by a postdoctoral fellowship of the German Academic Exchange Service (DAAD).

References

Ben Ambridge, Julian M. Pine, Caroline F. Rowland, Daniel Freudenthal, and Franklin Chang. 2014. Avoiding dative overgeneralisation errors: semantics, statistics or both? *Language, Cognition and Neuroscience*, 29(2):218–243.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403.

J.L. Austin. 1975. *How To Do Things With Words: The William James Lectures delivered at Harvard University in 1955*. Oxford University Press.

Gordon P. Baker and P. M. S. Hacker. 1986. *Language, sense and nonsense: a critical investigation into modern theories of language*, reprinted edition. Blackwell, Oxford.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Johan J. Bolhuis, Stephen Crain, and Ian Roberts. 2023. Language and learning: the cognitive revolution at 60-odd. *Biological Reviews*, 98(3):931–941.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Laurel Brehm, Pyeong Whan Cho, Paul Smolensky, and Matthew A. Goldrick. 2022. Pips: A parallel planning model of sentence production. *Cognitive Science*, 46(2):e13079.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R Harald Baayen. 2007. "Predicting the Dative Alternation". In Gerlof Bouma, Irene Krämer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. KNAW.

Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.

Paco Calvo and John Symons. 2014. *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. MIT Press.

Rudolf Carnap. 1937. *Logical Syntax of Language*, 1st edition. Routledge.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

Morten H Christiansen and Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205.

Herbert H. Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359.

Albert Coil and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ishita Dasgupta, Demi Guo, Samuel J. Gershman, and Noah D. Goodman. 2020. Analyzing machine-learned representations: A natural language case study. *Cognitive Science*, 44(12):e12925.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Vittoria Dentella, Fritz Günther, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2024. Testing ai on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14(1):28083.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David R. Dowty. 1979. *The Semantics of Aspectual Classes of Verbs in English*, pages 37–132. Springer Netherlands, Dordrecht.

John W. Du Bois. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25(3):359–410.

Ewa Dąbrowska. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27(1):1–23.

Jeffrey L. Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582.

Cyn X Fang, Edward Gibson, and Moshe Poliak. 2023. Individual difference in sentence preferences vs. sentence completion abilities.

Jerry Fodor and Ernie Lepore. 2002. Why compositionality won't go away: Reflections on horwich's 'deflationary'theory. *Meaning and representations, ed. Emma Borg. Oxford: Blackwell*.

Jerry Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Gottlob Frege. 1918. Der gedanke. eine logische untersuchung. *Beiträge zur Philosophie des deutschen Idealismus*.

Edward Gibson and Gregory Hickok. 1993. Sentence processing with empty categories. *Language and Cognitive Processes*, 8(2):147–161.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.

Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Adele E. Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

HP Grice. 1975. Logic and conversation. *Syntax and semantics*, 3.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29(2):261–290.

Kurt Gödel. 1931. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für Mathematik und Physik*, 38:173–198.

Robert Hawkins, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663, Online. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Valentin Hofmann, Leonie Weissweiler, David R. Mortensen, Hinrich Schütze, and Janet B. Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19):e2423232122.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language Models Align with Human Judgments on Key Grammatical Constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

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

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*. To appear.

Tom S Juzek. 2024. The syntactic acceptability dataset (preview): A resource for machine learning and linguistic analysis of English. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16113–16120, Torino, Italia. ELRA and ICCL.

Jerrold J. Katz. 1972. *Semantic Theory*. Harper & Row, New York.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR.

Andrew K. Lampinen, Arslan Chaudhry, Stephanie C. Y. Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James L. McClelland. 2025. On the generalization of language models from in-context learning and finetuning: a controlled study. *Preprint*, arXiv:2505.00661.

Douglas B. Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38.

Cara Su-Yi Leong and Tal Linzen. 2024. Testing learning hypotheses using neural networks by manipulating learning data. *Preprint*, arXiv:2407.04593.

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.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. SLOG: A structural generalization benchmark for semantic parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics.

Matthias Lindemann, Alexander Koller, and Ivan Titov. 2024. Strengthening structural inductive biases by pre-training to perform syntactic transformations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11558–11573, Miami, Florida, USA. Association for Computational Linguistics.

Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Gary Marcus. 1998. Rethinking eliminative connectionism. *Cognitive Psychology*, 37(3):243–282.

Gary Marcus. 2001. *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press.

James L McClelland. 2015. Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. *The handbook of language emergence*, pages 53–80.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kate McCurdy, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao. 2024. Toward compositional behavior in neural models: A survey of current views. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9323–9339, Miami, Florida, USA. Association for Computational Linguistics.

Laura A Michaelis. 2003. Headless constructions and coercion by construction. *Mismatch: Form-function incongruity and the architecture of grammar*, pages 259–310.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Marvin Minsky and Seymour Papert. 1969. An introduction to computational geometry. *Cambridge tiass.*, HIT 479.480:104.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 221–242. Springer Netherlands, Dordrecht.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2022. Probing for labeled dependency trees. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7711–7726, Dublin, Ireland. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Timothy J O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

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

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Barbara H. Partee. 1984. Nominal and temporal anaphora. *Linguistics and Philosophy*, 7(3):243–286.

Andrew Pawley and Frances Hodgetts Syder. 1983. Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics*, 7(5):551–579.

Steven T. Piantadosi. 2024. Modern language models refute chomsky's approach to language. In Edward Gibson and Moshe Poliak, editors, *From fieldwork to linguistic theory: A tribute to Dan Everett*. Language Science Press.

Martin J. Pickering and Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin*, 134(3):427–459.

Steven Pinker. 1999. *Words and Rules: The Ingredients of Language*. Basic Books, New York.

Steven Pinker and Alan Prince. 1988. On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. *Cognition*, 28(1-2):73–193.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Pawel Nowak, Tal Linzen, Fei Sha, and Kristina Toutanova. 2022. Improving compositional generalization with latent structure and data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4341–4362, Seattle, United States. Association for Computational Linguistics.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.

Clarice Robenalt and Adele E. Goldberg. 2015. Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics*, 26(3):467–503.

Timothy T. Rogers and James L. McClelland. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.

Hayley Ross, Kathryn Davidson, and Najoung Kim. 2024. Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 131–153, Miami, Florida, USA. Association for Computational Linguistics.

David E. Rumelhart, James L. McClelland, and PDP Research Group. 1986. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.

Jacob Russin, Sam Whitman McGrath, Danielle J. Williams, and Lotem Elber-Dorozko. 2024. From frege to chatgpt: Compositionality in language, cognition, and deep neural networks. *Preprint*, arXiv:2405.15164.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carson T Schütze and Jon Sprouse. 2013. Judgment data. *Research methods in linguistics*, pages 27–50.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.

Dan I Slobin. 1986. The acquisition and use of relative clauses in turkic and indo-european languages. *Studies in Turkish linguistics*, 8:273.

P. F. Strawson. 1967. Is existence never a predicate? *Crítica: Revista Hispanoamericana de Filosofía*, 1(1):5–19.

Leonard Talmy. 1975. Figure and ground in complex sentences. In *Annual meeting of the Berkeley linguistics society*, pages 419–430.

Alfred Tarski. 1944. The semantic conception of truth: and the foundations of semantics. *Philosophy and Phenomenological Research*, 4(3):341–376.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the*

28th International Conference on Computational Linguistics, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369, Marseille, France. European Language Resources Association.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

Johan van Benthem. 2008. A brief history of natural logic. In M. Chakraborty, B. Löwe, M. Nath Mitra, and S. Sarukki, editors, *Logic, Navya-Nyaya and Applications: Homage to Bimal Matilal*. College Publications.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Pia Weißenhorn, Lucia Donatelli, and Alexander Koller. 2022. Compositional generalization with a broad-coverage semantic parser. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 44–54, Seattle, Washington. Association for Computational Linguistics.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Terry Winograd. 1980. What does it mean to understand language? *Cognitive Science*, 4(3):209–241.

Ludwig Wittgenstein. 1953. *Philosophische Untersuchungen*. Suhrkamp Verlag, Frankfurt am Main.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions are so difficult that Even large language models get them right for the wrong reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811, Torino, Italia. ELRA and ICCL.

You Shall Know a Construction by the Company it Keeps: Computational Construction Grammar with Embeddings

Lara Verheyen
Artificial Intelligence Laboratory
Vrije Universiteit Brussel, Belgium
lara.verheyen@ai.vub.ac.be

Jonas Doumen
Faculty of Arts
KU Leuven, Belgium
jonas.doumen@kuleuven.be

Paul Van Eecke*
Artificial Intelligence Laboratory
Vrije Universiteit Brussel, Belgium
paul@ai.vub.ac.be

Katrien Beuls*
Faculté d'informatique
Université de Namur, Belgium
katrien.beuls@unamur.be

Abstract

Linguistic theories and models of natural language can be divided into two categories, depending on whether they represent and process linguistic information *numerically* or *symbolically*. Numerical representations, such as the embeddings that are at the core of today's large language models, have the advantage of being learnable from textual data, and of being robust and highly scalable. Symbolic representations, like the ones that are commonly used to formalise construction grammar theories, have the advantage of being compositional and interpretable, and of supporting sound logic reasoning. While both approaches build on very different mathematical frameworks, there is no reason to believe that they are incompatible. In the present paper, we explore how numerical, in casu distributional, representations of linguistic forms, constructional slots and grammatical categories can be integrated in a computational construction grammar framework, with the goal of reaping the benefits of both symbolic and numerical methods.[1]

1 Introduction

Linguistic theories and models of natural language typically fall into one of two categories. The first category represents and processes linguistic information *symbolically*, adopting formal logic as the underlying framework. The second category represents and processes linguistic information *numerically*, adopting the framework of linear algebra.

The symbolic approach is widely used to formalise construction grammar theories (Fillmore, 1988; Kay and Fillmore, 1999; Steels and De Beule, 2006; Michaelis, 2008; Sag, 2012), with symbolic programming techniques forming the backbone of their computational implementations (Bergen and Chang, 2005; Steels and De Beule, 2006; van Trijp et al., 2022). Symbolic representations bring the advantage of being compositional and interpretable, and of supporting sound logic reasoning.

The numerical approach is widely adopted in the field of natural language processing (NLP), and lies for example at the core of today's large language models (LLMs) (Mikolov et al., 2013; Vaswani et al., 2017; Lenci, 2018; Devlin et al., 2019). In essence, numerical representations of linguistic information are learnt from textual data, thus based on the distribution of tokens with respect to each other. Apart from being learnable from raw textual input, distributional representations bring the advantage of being robust against noise, of generalising well to new data, and of scaling effectively with respect to growing amounts of input data from different domains.

As both approaches are rooted in very different mathematical frameworks, namely formal logic versus linear algebra, the integration of concepts and techniques from both fields is not straightforward. At the same time, logic-based and distributional approaches are widely regarded as complementary, and there exists no a priori reason to believe that they would be in any way incompatible.

In this paper, we explore how distributional representations can be integrated in a computational construction grammar framework, and how this integration of symbolic and numerical methods can enhance the robustness and generality of constructional language processing. In particular, we show how distributional representations of (i) linguistic forms, (ii) constructional slots, and (iii) gram-

*Joint last authors.

[1]The authors declare that this paper was conceived and written without the assistance of generative writing aids.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 75–83

matical categories can be integrated into the data structures and algorithms that underlie Fluid Construction Grammar (FCG) (Steels, 2004; Beuls and Van Eecke, 2023). Through a variety of examples, we demonstrate how this integration can benefit a broad-coverage FCG grammar learnt from PropBank-annotated corpora. Finally, we conclude that the future of construction grammar is neither symbolic nor numerical, but lies in a combination of both paradigms.

2 Background

For the purposes of this exploration, we start from a symbolic construction grammar that was learnt from a collection of corpora in which English utterances were semantically annotated with PropBank rolesets (Palmer et al., 2005).[2] The grammar was learnt using the Fluid Construction Grammar framework (Beuls and Van Eecke, 2023) and holds 21,052 constructions that can be used to annotate open-domain English utterances with argument structure information in the form of semantic frames.

The basic architecture of the grammar is laid out in Figure 1. The input to the grammar consists of an utterance, in this case *"The doctor wrote him a prescription."*, which is analysed on the fly into its immediate constituents using the Berkeley neural parser (Stern et al., 2017) (see Step ①). A first type of construction identifies possible frame-evoking elements based on their lemma and part-of-speech tag. Here, the WRITE(V)-CXN indicates that the constituent 'unit-4' might represent a frame-evoking element by adding the for now underspecified `roleset` feature to this unit, along with a lexical category proper to the WRITE(V)-CXN (see Step ②). The resulting unit is shown as 'unit-4a'. The addition of a lexical category unlocks the application of a second type of construction that attributes semantic roles based on an utterance's constituent layout. In the example, a ditransitive construction that is compatible with the category contributed by the WRITE(V)-CXN respectively attributes the roles 'arg0' (prototypical agent),'arg1' (prototypical patient) and 'arg2' (prototypical beneficiary) to the constituents 'unit-2', 'unit-6' and 'unit-5'. The ditransitive construction also contributes its own grammatical category to the unit containing the frame-evoking element (see Step ③). The result is shown as 'unit-4b'. A final construction that is compatible with both the lexical category contributed by the WRITE(V)-CXN and the grammatical category contributed by the ditransitive construction fills out the value of the `roleset` feature, in this case PropBank's `write.01` roleset (see Step ④). The result is shown as 'unit-4c'. As the example utterance only expresses a single frame, the construction application process stops here. The resulting frame is then extracted and rendered into a more human-readable format (see Step ⑤). All constructions as well as the categorial links that express compatibility between constructional units were learned from corpus data using FCG's `fcg-propbank` subsystem (Van Eecke and Beuls, 2025).

[2]In particular, the examples throughout this paper were selected from the test sets of the OntoNotes (Weischedel et al., 2013) and English Web Treebank (EWT) (Bies et al., 2012) corpora, while the grammar itself was learnt from the training sets of the same corpora.

3 Distributional Representations of Linguistic Forms

A classical argument against symbolic methods revolves around their reliance on exact matches between symbols. For example, the symbol DOG is in its representation not any more closely related to the symbol PUPPY than it is to the symbols CAT or PHILOLOGY. Representationally, symbols are either equal to or different from each other. Standard FCG builds on this property for implementing the process of construction application, where features and values in the pre- and postconditions of constructions are unified with their counterparts in the transient structure based on the equality of symbols (Steels and De Beule, 2006). The classical argument against symbolic methods points to the brittleness of relying on exact matches, as symbolic models tend to have difficulties handling input that even slightly deviates from what is expected. Distributional methods on the other hand represent linguistic forms in a vector space, where forms are compared in terms of distributional similarity rather than representational equality. In such models, the distance between DOG and PUPPY will effectively be smaller than the distance between DOG and PHILOLOGY.

Take for example the utterance "*So I mean that right there it enraged me.*" (OntoNotes `bc/cnn/00/cnn_0000`), which expresses an instance of the `mean.01` roleset and an instance of the `enrage.01` roleset. The base grammar from the previous section however, only retrieves the

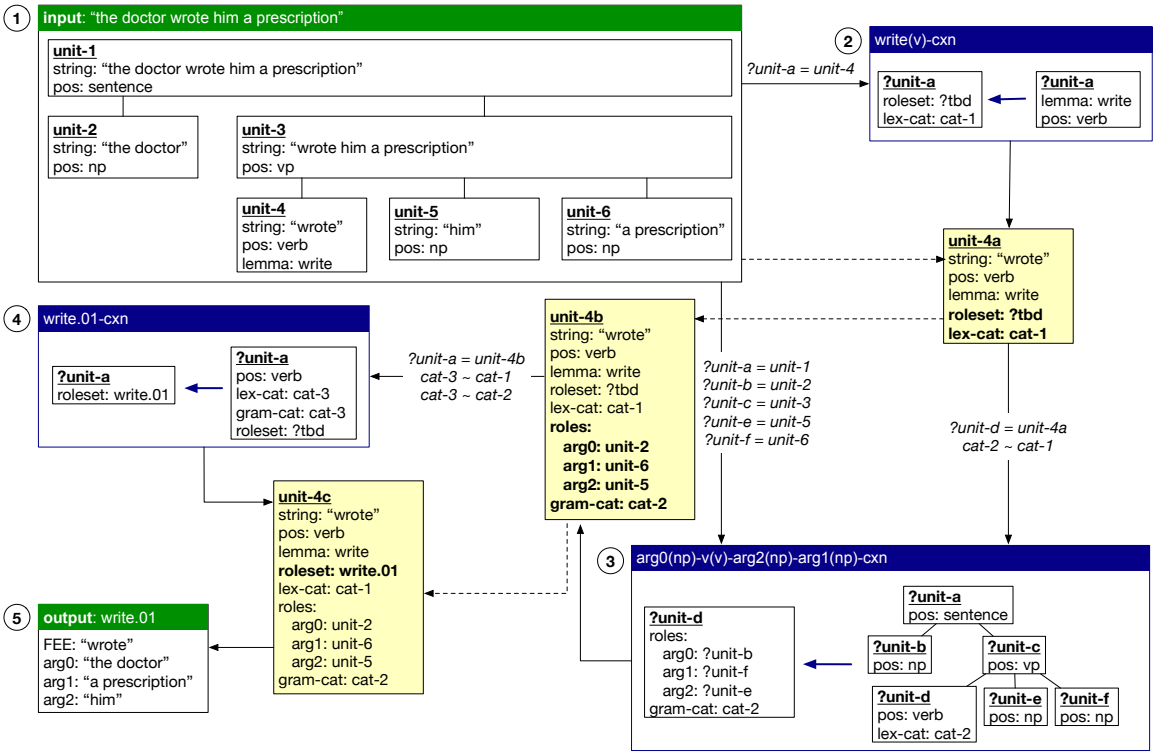


Figure 1: Illustrative example of the symbolic base grammar comprehending "*The doctor wrote him a prescription.*" ①. The WRITE(V)-CXN identifies a potential frame-evoking element ②. A ditransitive construction then attributes the semantic roles of agent ('arg0'), patient ('arg1') and beneficiary ('arg2') to particular constituents ③. The WRITE.01-CXN determines the roleset (`write.01`) of the evoked frame ④, after which the result is shown ⑤.

instance of the `mean.01` roleset. Upon closer inspection, it turns out that the verb "*enrage*" did not occur anywhere in the training corpus and that consequently no construction was learnt that identifies "*enraged*" as a possible frame-evoking element. At the same time, many constructions were learnt for other verbs that are distributionally close to "*enrage*" (such as "*anger*", "*madden*" or "*infuriate*") and that even appear in similar argument structure constructions ("*NP:Arg0 (angers — maddens — infuriates) NP:Arg1*"). The reason why these constructions cannot apply is simply that there is no exact match between the lemma of the observed token ("*enrage*") and the lemmas incorporated in the constructions ("*anger*", "*madden*" and "*infuriate*").

As a first step in the integration of symbolic and distributional methods, we will represent lemmata distributionally rather than symbolically in FCG constructions and transient structures. Concretely, we substitute the `lemma` features in the units of the input transient structure by `embedding` features that hold as their value pointers to pre-trained, 100-dimensional GloVe embeddings of the original lemmata (Pennington et al., 2014). This is shown for the example utterance "*So I mean that right there it enraged me.*" in Step ① of Figure 2. Likewise, we substitute the `lemma` features in the constructions of the grammar by `embedding` features that point to pre-trained GloVe embeddings (see Step ②). The INFURIATE(V)-CXN thereby does not match on the symbol INFURIATE any more but on the GloVe embedding for the form "*infuriate*". We also modify FCG's unification algorithms in such a way that they no longer compute symbol equality when handling vectors, but compute their cosine similarity. These similarities are then used to create scored unification results and rank possible construction applications. In the example, the highest-ranked result is yielded by the INFURIATE(V)-CXN, which matches on the unit holding "*enraged*" with a cosine similarity of 0.84. Then, the transitive construction that was learned during training to be compatible with the INFURIATE(V)-CXN can apply, followed by the INFURIATE.01-CXN. This results in the extraction of an instance of the `infuriate.01` roleset, with "*enraged*" as the frame-evoking element and "*it*" and "*me*" respectively as its 'arg0' ('*causer of anger*') and 'arg1' ('*angry entity*').

This example demonstrates how constructions

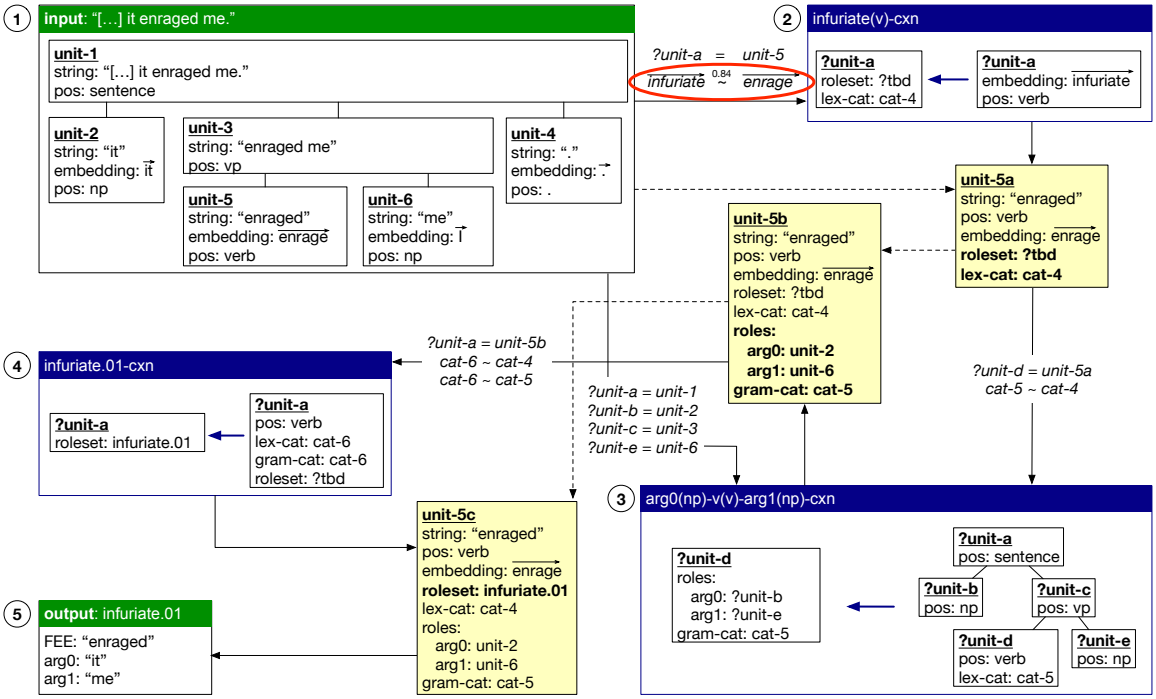


Figure 2: Schematic illustration of the integration of distributional token representations in constructional language processing. The INFURIATE(V)-CXN identifies "*enraged*" as a possible frame-evoking element based on the high cosine similarity between the embeddings for "*enrage*" and "*infuriate*", recovering from the absence of the token "*enrage*" in the training corpus.

can apply without requiring a perfect symbolic match, relying on the distributional closeness of forms, in this case the lemmata of potential frame-evoking elements. This was achieved by integrating numerical representations of linguistic information (i.c. word embeddings) and operations over them (i.c. cosine computation) with symbolic representations (i.c. feature structures) and operations over these (i.c. unification). In fact, this integration can be considered an extension of the way matches between categories in the categorial network of a grammar were already integrated into FCG's unification algorithms (see Van Eecke, 2018).

4 Distributional Representations of Constructional Slots

Now that we have represented the substantive material in constructions, such as word forms and lemmata, using word embeddings, we take the same idea a step further and integrate distributional representations of constructional slots. Let us consider as an example the utterance "*Jesus taught the people in the Temple area every day.*" (OntoNotes `ontonotes/pt/nt/42/nt_4219`). The base grammar yields two competing analyses which it considers equally fit. Both analyses identify an instance of the `teach.01` roleset, in which "*Jesus*" takes up the role of 'arg0' ('*teacher*'). One analysis assigns the role of 'arg1' ('*subject*') to "*the people*", while the other assigns it the role of 'arg2' ('*student(s)*'). The two analyses differ in the argument structure construction that is used. In the first analysis, a transitive construction applies that maps the noun phrase after the verb to the 'arg1' role, whereas in the second analysis, a construction applies that maps this noun phrase to the 'arg2' role. Both constructions can be traced back to utterances in the training corpus, such as "*Her mother taught [Sunday School]$_{arg1}$ for 50 years.*" (OntoNotes `bn/cnn/03/cnn_0324`) and "*You teach [others]$_{arg2}$, so why don't you teach [yourself]$_{arg2}$?*" (OntoNotes `pt/nt/45/nt_4502`). This ambiguity cannot be resolved on the level of the morphosyntactic structure of the utterances and necessitates modelling the lexical content of the slot fillers.

We extend the idea of including an `embedding` feature to the units in the initial transient structure also to phrasal units. The embeddings on phrasal level are in this prototype computed as the sum of the GloVe embeddings of the lemmas of their constituent parts (see Step ① in Figure 3). In each

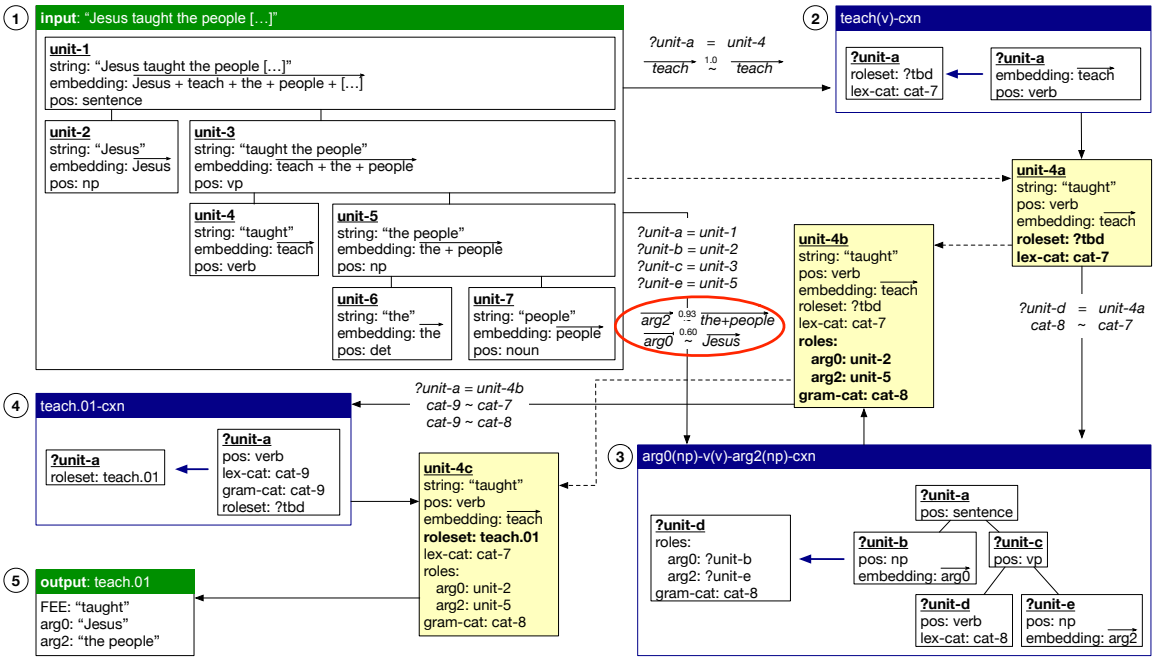


Figure 3: Schematic illustration of the integration of distributional information for representing prototypical slot fillers within argument structure constructions. The embeddings in the argument structure constructions are computed based on their fillers as observed in the training corpus.

argument structure construction, we also add an `embedding` feature to all units that are assigned a role (see Step ③). These embeddings are computed by averaging the summed embeddings of all lemmata for all fillers observed in a particular slot during training. For example, the value of the `embedding` feature in the 'arg1' slot of a transitive construction would point to a vector representing the prototypical patient/undergoer that fills that slot. The unification algorithm described in the previous section, which computes cosine similarities when handling vectors, is again used. In our example, this leads to two construction application results, one for each of the two argument structure constructions, with the one where the 'arg2' role is taken up by "*the people*" is ranked highest. Indeed, the match between "*the people*" and the prototypical vector of the 'arg2' slot of this construction is considerably higher than the match between "*the people*" and the prototypical vector for the 'arg1' slot in the other construction. The highest-ranked solution thereby yields a correct semantic role assignment.

While the previous section and the current section have both integrated distributional representations into FCG constructions, the impact on the grammar is quite different. In the previous section, symbols representing substantive material in constructions were substituted by pointers to embeddings. This has rendered the constructions more general and less specific to particular input structures, as exact matches between symbols are no longer a hard constraint. In the present section, the embeddings were introduced to represent the prototypical lexical content of constructional slots and do not replace a feature that was present in the base grammar. The constructions have thereby become more specific, allowing for a more fine-grained disambiguation between possible construction application results. The integration of embeddings should thus not be seen solely as a means to make symbolic grammars more general, but it can also serve to integrate more specific information into constructions that would be considered too specific when relying on exact matching.

5 Distributional Representations of Grammatical Categories

In the previous sections, we have integrated pretrained GloVe embeddings in the base grammar to distributionally represent linguistic forms and prototypical slot fillers. These embeddings were trained independently from the base grammar on large amounts of text and mainly reflect the lexical content of words and phrases. In this section, we explore a different approach to integrating dis-

tributional representations in constructions. We no longer rely on externally trained embeddings, but model the similarity between grammatical categories based on the constructional slots they are compatible with. A weighted graph capturing the frequency of these slot-filler relations is built up while the grammar is being learnt from corpus data.

Let us consider the example utterance "*Try googling it for more info.*" (English Web Treebank `answers/00/20080426141111AAgPUwU_ans`). The base grammar identifies "*googling*" as a potential frame-evoking element, but holds no argument structure construction that is both compatible with the lemma `google` and the imperative transitive structure in which it appears syntactically. Consequently, no instance of the `google.01` roleset is being detected using the base grammar and no semantic roles are assigned. Importantly, the reason is not that the imperative transitive construction was not learnt during training, but that it was not learnt to be compatible with the category proper to the GOOGLE(V)-CXN.

Based on the weighted graph that captures the distribution of slot-filler categories over constructional slots, similarity between categories can be computed using the weighted cosine similarity metric. As such, slot-filler categories that are similarly distributed over constructional slots will be closer to each other than categories that rarely occur in the same constructions. In the base grammar, the category proper to the GOOGLE(V)-CXN bears a high similarity to the category proper to the DISREGARD(V)-CXN. Intuitively, this is not surprising, as both verbs are strictly transitive. If the distributions of two categories are close to each other, which means that the two categories behave similarly in the grammar, one could infer that if one category is compatible with a specific constructional slot, the other category is also likely to be compatible with it. In our example, the compatibility of the category proper to the DISREGARD(V)-CXN with the category matched by the frame-evoking element unit of the imperative transitive construction can be taken as an indication that this specific argument structure construction might also provide a correct role assignment for the GOOGLE(V)-CXN. Indeed, the imperative transitive construction here correctly assigns the 'arg1' role ('*target of search*') to "*it*". The processing of this example utterance is schematically depicted in Figure 4. The link in the categorial network between `cat-10` (GOOGLE(V)-CXN) and `cat-11` (V(V)-ARG1(NP)-CXN), which is necessary to apply the imperative transitive construction is inferred on the fly with a graph cosine similarity score of 0.3 based on the distributional similarity between `cat-10` (GOOGLE(V)-CXN) and `cat-21` (DISREGARD(V)-CXN).

6 Related Work

While we provide to the best of our knowledge the first fully operational and computationally implemented prototype of a symbolic construction grammar that integrates distributional representations and processing mechanisms to enhance its robustness and generality, many scholars have already addressed in one way or another the challenge of combining construction grammar with distributional semantics. Levshina and Heylen (2014) pioneered the use of distributional representations to represent the prototypical slot-fillers of constructions in a corpus-linguistic study. Hilpert and Perek (2015) and Perek (2016) have used distributional representations to track changes in the slot-fillers of constructions over time. In the same spirit, Lebani and Lenci (2018) make use of distributional representations to represent thematic roles. Rambelli et al. (2019) and Blache et al. (2024) make a case for integrating distributional representations into construction grammar and present a theoretical proposal of how distributional representations could be integrated into Sign-Based Construction Grammar to represent word forms and slots. Finally, Dunn (2017, 2024) provides a grammar induction algorithm that makes use of distributional representations to model the prototypical content of constructional slots. A related body of research is not directly concerned with construction grammar, but with the integration of formal and distributional semantics (for an overview, see Boleda and Herbelot, 2016, and other papers in the same special issue). The goal is again to combine the compositional and inferential aspects of logic-based representations with the machine learnability and lexical modelling capacities of distributional representations.

A more distantly related line of research that is concerned with both construction grammar and word embeddings investigates the linguistic capabilities of large language models from a construction grammar perspective. The goal is not to integrate symbolic and distributional approaches, but to assess to what extent distributional approaches, in

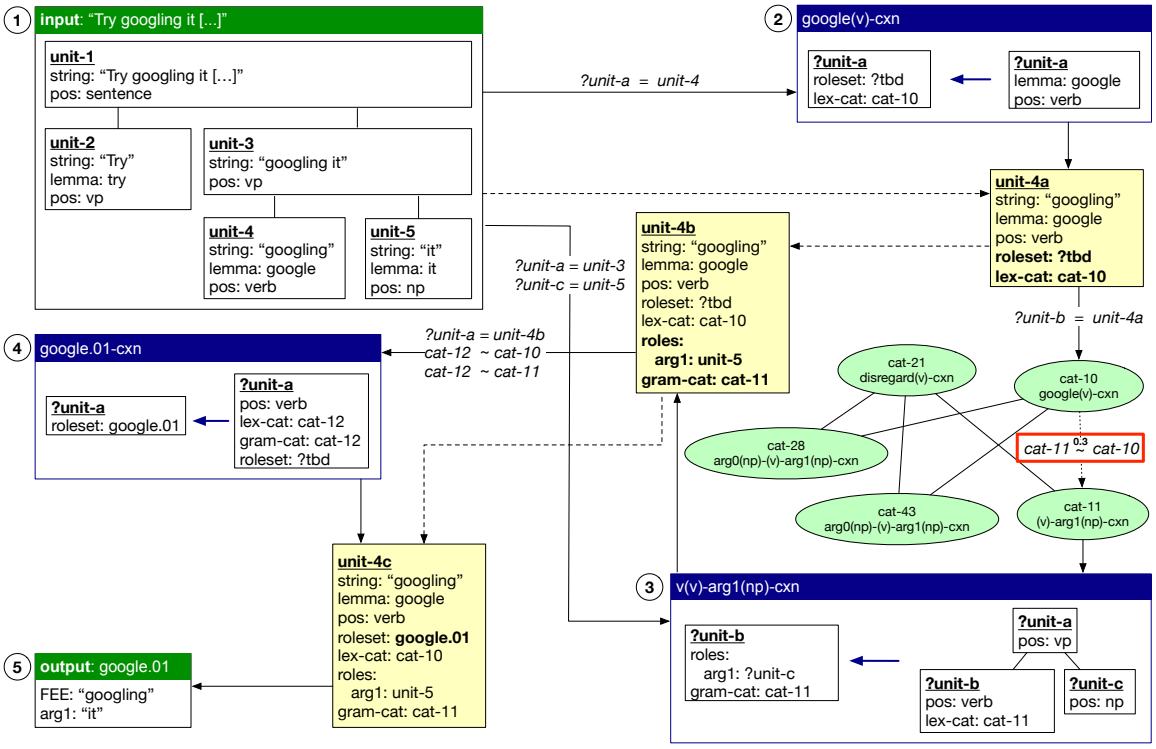


Figure 4: Schematic illustration of the integration of distributional representations of grammatical categories. The category that is proper to the GOOGLE(V)-CXN is not directly compatible with the category of the frame-evoking element slot of the imperative transitive construction (see ③). However, this categorial link is inferred on the fly based on the close distributional similarity between `cat-10` and `cat-21`.

particular large language models, capture the constructional knowledge that is typically represented symbolically in the construction grammar literature (see e.g. Tayyar Madabushi et al., 2020; Weissweiler et al., 2022, 2023; Bonial and Tayyar Madabushi, 2024; Zhou et al., 2024; Tayyar Madabushi et al., 2025).

7 Discussion and Conclusion

We have started from the observation that linguistic theories and models of natural language typically adopt either a symbolic or a numerical approach. At the same time, symbolic and numerical approaches are widely acknowledged to be complimentary to each other (see e.g. Boleda and Herbelot, 2016). Symbolic approaches have the advantage of supporting compositionality, interpretability and sound logic inference, whereas numerical approaches have the advantage of being more scalable, robust and easier to learn from data. The integration of symbolic and numerical approaches is however complicated by the fact that they are rooted in very different mathematical frameworks, namely formal logic versus linear algebra.

In this paper, we have explored the integration of numerical representations, in this case distributional representations of word forms, constructional slots and grammatical categories, in a symbolic computational construction grammar framework. Concretely, we have shown how such representations can be operationalised in Fluid Construction Grammar and enhance the robustness and generality of learned FCG grammars. In a first experiment, we have replaced the substantive material in the constructions of a learned, symbolic base grammar by pre-trained GloVe embeddings of the same material. By extending FCG's unification algorithms to compute cosine similarities instead of symbol equalities during the construction application process, we obtained a range of ranked construction application results in cases where there was no exact match, but a close match, between the lemma required by a construction and the one observed in the input utterance. In a second experiment, we have integrated vector representations of the prototypical lexical content of constructional slots to aid disambiguation where competing constructions could apply. The vectors were computed while the grammar was being learned, based on pre-trained GloVe embeddings of the words and

phrases that were observed in the respective slots of the construction. By aggregating the cosine similarities of slots and their fillers during construction application, we again obtained a range of construction application results ranked according to their lexical fit with the applied constructions. In a third experiment, we no longer relied on externally trained embeddings, but have modelled the similarity between grammatical categories based on their observed distribution over constructional slots. This distribution was then used to create links on the fly in the categorial network that were never learnt during training.

The experiences gained while working on this initial prototype have convinced us that the future of computational construction grammar will be hybrid. Yet, further research is now needed to scale this prototype for large-scale evaluation, where the advantages of integrating distributional representations can also be shown quantitatively.

Acknowledgements

The research reported on in this paper was funded by the F.R.S.-FNRS-FWO WEAVE project HERMES I under grant numbers T002724F (F.R.S.-FNRS) and G0AGU24N (FWO), the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme, and the AI Flagship project ARIAC by DigitalWallonia4.ai.

References

Benjamin K. Bergen and Nancy Chang. 2005. Embodied Construction Grammar in simulation-based language understanding. In Mirjam Fried and Jan-Ola Östman, editors, *Construction Grammars: Cognitive Grounding and Theoretical Extensions*, pages 147–190. John Benjamins, Amsterdam, Netherlands.

Katrien Beuls and Paul Van Eecke. 2023. Fluid Construction Grammar: State of the art and future outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 41–50. Association for Computational Linguistics.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank ldc2012t13. Philadelphia, Linguistic Data Consortium.

Philippe Blache, Emmanuele Chersoni, Giulia Rambelli, and Alessandro Lenci. 2024. Composing or Not Composing? Towards Distributional Construction Grammars. *arXiv preprint arXiv:2412.07419*.

Gemma Boleda and Aurélie Herbelot. 2016. Formal Distributional Semantics: Introduction to the Special Issue. *Computational Linguistics*, 42(4):619–635.

Claire Bonial and Harish Tayyar Madabushi. 2024. Constructing understanding: on the constructional information encoded in large language models. *Language Resources and Evaluation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis.

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

Jonathan Dunn. 2024. *Computational construction grammar: A usage-based approach*. Cambridge University Press.

Charles J. Fillmore. 1988. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 35–55.

Martin Hilpert and Florent Perek. 2015. Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, 1(1):339–350.

Paul Kay and Charles Fillmore. 1999. Grammatical constructions and linguistic generalizations: The what's X doing Y? construction. *Language*, 75(1):1–33.

Gianluca Lebani and Alessandro Lenci. 2018. A distributional model of verb-specific semantic roles inferences. In Thierry Poibeau and Aline Villavicencio, editors, *Language, cognition, and computational models*, pages 118–158. Cambridge University Press.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4(1):151–171.

Natalia Levshina and Kris Heylen. 2014. A radically data-driven construction grammar: Experiments with dutch causative constructions. In Ronny Boogaart, Timothy Colleman, and Gijsbert Rutten, editors, *Extending the Scope of Construction Grammar*, pages 17–46. De Gruyter Mouton, Berlin, Germany.

Laura A. Michaelis. 2008. Entity and event coercion in a symbolic theory of syntax. In Jan-Ola Östman and Mirjam Fried, editors, *Construction grammars: Cognitive grounding and theoretical extensions*, pages 45–88. John Benjamins Publishing Company.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 1–9, Red Hook.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Florent Perek. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1):149–188.

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)*.

Ivan A. Sag. 2012. Sign-based construction grammar: An informal synopsis. In Hans C. Boas and Ivan A. Sag, editors, *Sign-based construction grammar*, pages 69–202. CSLI Publications/Center for the Study of Language and Information, Stanford, CA, USA.

Luc Steels. 2004. Constructivist development of grounded construction grammar. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 9–16. Association for Computational Linguistics.

Luc Steels and Joachim De Beule. 2006. Unify and merge in Fluid Construction Grammar. In *Symbol Grounding and Beyond, Third International Workshop on the Emergence and Evolution of Linguistic Communication, EELC 2006, Rome, Italy, September 30 - October 1, 2006, Proceedings*, volume 4211 of *Lecture Notes in Computer Science*, pages 197–223. Springer.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032. International Committee on Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Petar Milin, and Dagmar Divjak. 2025. Construction grammar and language models. In Mirjam Fried and Kiki Nikiforidou, editors, *The Cambridge Handbook of Construction Grammar*, pages 572–595. Cambridge University Press, Cambridge, United Kingdom.

Paul Van Eecke. 2018. *Generalisation and specialisation operators for computational construction grammar and their application in evolutionary linguistics Research*. Ph.D. thesis, Vrije Universiteit Brussel, Brussels: VUB Press.

Paul Van Eecke and Katrien Beuls. 2025. PyFCG: Fluid Construction Grammar in Python. *arXiv preprint arXiv:2505.12920*.

Remi van Trijp, Katrien Beuls, and Paul Van Eecke. 2022. The FCG Editor: An innovative environment for engineering computational construction grammars. *PLOS ONE*, 17(6):e0269708.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6000–6010, Long Beach.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 ldc2013t19. Philadelphia, Linguistic Data Consortium.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95. Association for Computational Linguistics.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882. Association for Computational Linguistics.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions are so difficult that even large language models get them right for the wrong reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811. Association for Computational Linguistics.

Constructions All the Way Up: From Sensory Experiences to Construction Grammars

Jérôme Botoko Ekila*
Artificial Intelligence Laboratory
Vrije Universiteit Brussel, Belgium
jerome@ai.vub.ac.be

Lara Verheyen*
Artificial Intelligence Laboratory
Vrije Universiteit Brussel, Belgium
lara.verheyen@ai.vub.ac.be

Katrien Beuls†
Faculté d'informatique
Université de Namur, Belgium
katrien.beuls@unamur.be

Paul Van Eecke†
Artificial Intelligence Laboratory
Vrije Universiteit Brussel, Belgium
paul@ai.vub.ac.be

Abstract

Constructionist approaches to language posit that all linguistic knowledge is captured in constructions. These constructions pair form and meaning at varying levels of abstraction, ranging from purely substantive to fully abstract and are all acquired through situated communicative interactions. In this paper we provide computational support for these foundational principles. We present a model that enables an artificial learner agent to acquire a construction grammar directly from its sensory experience. The grammar is built from the ground up, i.e. without a given lexicon, predefined categories or ontology and covers a range of constructions, spanning from purely substantive to partially schematic. Our approach integrates two previously separate but related experiments, allowing the learner to incrementally build a linguistic inventory that solves a question-answering task in a synthetic environment. These findings demonstrate that linguistic knowledge at different levels can be mechanistically acquired from experience.

1 Introduction

According to constructionist approaches to language (Fillmore, 1988; Goldberg, 1995; Croft, 2001; Goldberg, 2003) all linguistic knowledge is captured in constructions, pairing form and meaning. Within this framework, constructions vary in their level of abstraction, ranging from purely substantive to fully abstract, all shaped by usage. As Goldberg (2003, p. 223) famously put it: *"it's constructions all the way down"*.

Constructions are not abstract templates shared uniformly between members of a linguistic community, rather each one is grounded in an individual's embodied experience and interaction with the world (Lakoff, 1987; Langacker, 1987; Bybee, 2010; Tomasello, 2003; Diessel, 2017). For instance, a construction mapping the form *"dog"* to its underlying DOG concept is shaped by an individual's encounters with dogs, including what they have seen, learned or heard about them. Beyond the perceptual level, language users also acquire constructions that coordinate more abstract cognitive processes (Goldberg, 1995). Consider the sentence *"The dog chases the cat."* in which the transitive construction organises the relation between a CHASING event and its participants. This abstract relation is learned through repeated encounters of linguistic utterances and observations in the world. Whether the meaning of a construction is a concept grounded in direct sensory experience or an abstract schema, all are pairings of form and meaning and arise from situated interactions (Beuls and Van Eecke, 2025). This linguistic knowledge is built up through cognitive mechanisms that reconstruct the intended meaning of an interlocutor and find patterns over form-meaning mappings (Tomasello, 2003; Dąbrowska and Lieven, 2005; Behrens, 2009; Lieven, 2014).

A computational approach to modelling language acquisition involves language games, in which embodied agents acquire constructions through repeated situated communicative interactions (Steels, 1995, 1999). These simulations offer a mechanistic model of language acquisition, and have been used to study the emergence of linguistic structure at multiple levels, from basic grounded

*Joint first authors.
†Joint last authors.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 84–95
September 24, 2025,

lexicons (Steels, 1995; Kaplan et al., 1998; Loetzsch, 2015; Nevens et al., 2020; Botoko Ekila et al., 2024) to early forms of syntax (De Beule and Bergen, 2006; Van Eecke, 2018) and more complex grammatical systems (van Trijp, 2008; Beuls and Höfer, 2011; Spranger and Steels, 2015; Steels and Garcia Casademont, 2015; Nevens et al., 2022; Doumen et al., 2024). However, a key challenge remains unsolved: no existing computational model has yet demonstrated the emergence of a construction grammar that is both directly learned from sensory experience and capable of capturing a range of constructions, spanning from fully substantive constructions to more abstract constructions, without a given lexicon, ontology or predefined categories.

In this paper, we present a model that enables a learner agent to acquire a construction grammar from the ground up through situated communicative interactions with a tutor agent. Using a curriculum learning approach, where training progresses from simpler to more complex interactions, the learner develops a grammar that spans from perceptually grounded lexical constructions to partially schematic constructions. We validate our approach experimentally in a synthetic continuous environment in which a learner develops a grammar to interpret and answer questions. We thereby demonstrate that, with the help of a tutor, a computational construction grammar including more abstract constructions can be acquired directly from sensory experience, supporting the hypothesis that it is also, indeed, *constructions all the way up*.

2 Background

The model we present is embedded within the framework of language games (Steels, 1995, 1999), which is used to simulate how agents can establish linguistic conventions through repeated situated communicative interactions. In this paper, we build on language acquisition experiments that each focus on different levels of abstraction: (i) acquiring perceptually grounded lexical constructions that link sensory experiences to linguistic forms (Nevens et al., 2020; Botoko Ekila et al., 2024) and (ii) acquiring grammatical constructions that capture structural patterns in language use (Nevens et al., 2022; Doumen et al., 2024). Although these four experiments focus on acquiring constructions at varying levels of abstraction, they rely on the same shared principle: agents acquire form-meaning mappings through situated communicative interactions (Beuls and Van Eecke, 2024). The next sections summarise the core mechanisms behind each experiment, which forms the basis of our integrated approach.

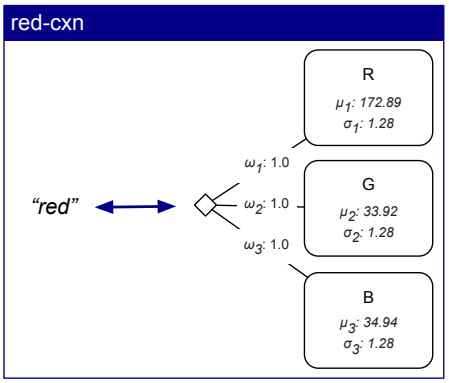


Figure 1: Example of a grounded lexical construction learned by the learner agent for the word *"red"*, which has specialised towards the three colour feature dimensions (RGB). Only dimensions with weights greater than 0.0 are shown.

2.1 Acquiring grounded lexical constructions

The experiments of Nevens et al. (2020) and Botoko Ekila et al. (2024) are concerned with acquiring form-meaning pairings that link sensory experiences to linguistic forms. In this process, a learner agent acquires a set of constructions that capture perceptual concepts such as RED or LARGE by interacting with a tutor agent. Importantly, the learner starts without any prior linguistic knowledge.

In the experiments, both agents are situated in a shared environment with different objects and engage in a series of referential games, each corresponding to a single interaction. In each interaction, the tutor (i) selects a target object from the scene and (ii) produces a single-word utterance that refers to a property of the selected object that distinguishes it from the other objects. The learner observes the scene through its own sensors, which capture raw perceptual features (e.g. RGB for colour or the number of pixels an object occupies in the image for size). The goal of the learner is to infer which object the tutor is referring to, based on the utterance, the perceptual input, and any linguistic knowledge acquired in previous interactions. After each interaction, the tutor reveals the correct referent (i.e. the target object), providing explicit feedback. At no point are the tutor's and learner's internal representations shared between agents. The learner must refine its own internal

representations through these interactions with the tutor. They store the observed word forms (e.g. *"red"*) and associated internal concept representations as form-meaning mappings in its inventory.

Concepts are modelled as weighted Gaussian distributions over sensory features. Each distribution captures the prototypical range of values associated with that feature, while the associated weight captures the feature's relevance to the concept. For example, as seen in Figure 1, the concept linked to the word *"red"* assigns high weights to RGB features and low weights to other features. These distributions and weights are updated incrementally through repeated interactions with the tutor.

Early on, the learner's answers are mostly incorrect, but as they interact more, the learner refines its concept representations based on the feedback of the tutor. Over time, the learner builds a conceptual system grounded in its own sensory experience of the world.

2.2 Acquiring grammatical constructions

In the experiments of Nevens et al. (2022) and Doumen et al. (2024), a learner acquires lexical and grammatical constructions by playing a question-answering game. The game operates in a symbolic representation of the environment of the experiments discussed in Section 2.1. In this symbolic version of the setting, objects are described using structured attribute-value pairs (e.g. OBJECT-1: {COLOUR: RED, SHAPE: CUBE}). This setup abstracts away from raw sensory inputs and perceptual processing, allowing the learner to work directly with high-level representations of objects. Thus, as seen in Figure 2, the meaning of the CUBES-CXN is represented by the symbol CUBE.

Within this symbolic setting, the tutor poses questions about a scene such as *"How many red cubes are there?"* or *"What shape does the blue object have?"*. The learner's task is to interpret the question and produce a correct answer. To achieve this, the learner builds a construction grammar that maps linguistic utterances to meaning representations that can be executed to retrieve the answer. To acquire these constructions, the learner is equipped with two core learning mechanisms: *intention reading* and *pattern finding* (Tomasello, 2003). Intention reading refers to a language user's ability to reconstruct the intended meaning of an utterance, enabling the learner to hypothesise about the speaker's intended meaning. Pattern finding

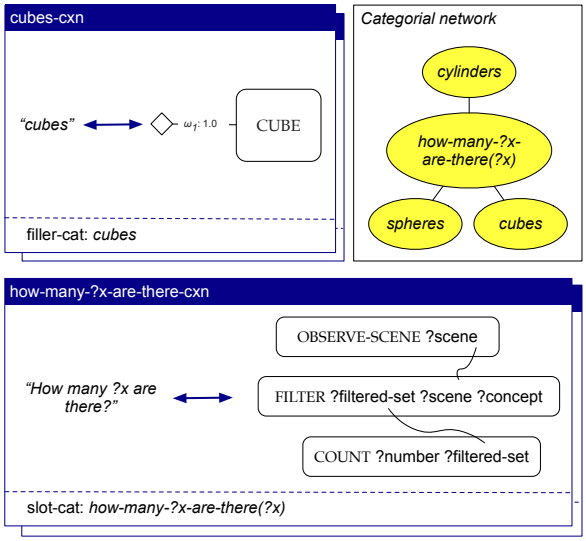


Figure 2: Example of two constructions acquired by the learner agent during the question-answering game that takes place in a symbolic environment. A lexical CUBES-CXN with a symbolic concept representation, an item-based HOW-MANY-?X-ARE-THERE-CXN and part of the categorial network, capturing the slot-filler relation through the categories in the constructions, are shown. Figure based on Nevens et al. (2022) and Doumen et al. (2024).

refers to the ability to generalise across different communicative interactions. We briefly summarise how these processes are operationalised, but for a more comprehensive explanation we refer the reader to Nevens et al. (2022) and Doumen et al. (2024).

The learner starts the game with an empty linguistic inventory but is endowed with a set of atomic cognitive operations (so-called *primitive operations*). The meaning of questions is represented as sequences of these operations, each of which are needed to find the correct answer, i.e. a form of procedural semantics (Winograd, 1972; Woods, 1968). Formally, each question is encoded as a set of predicates. Each predicate corresponds to a *primitive operation* that the learner can perform, such as filtering objects by their properties or counting elements in a set. For example, the question *"How many cubes are there?"* can be represented as a sequence of three primitive operations: (i) observing the current scene with OBSERVE-SCENE, (ii) filtering for objects of type cube with FILTER, and (iii) counting the resulting set with COUNT.

At the start of each interaction, both agents are situated in the same scene. The tutor then poses a question to the learner about the scene. The learner

attempts to interpret and answer the question using its current linguistic inventory. If the learner fails to interpret the question or the answer is incorrect, the tutor provides feedback in the form of the correct answer. The learner then attempts to recover the intended meaning by abductively reasoning about the tutor's communicative goal (i.e. *intention reading*). In doing so, it searches for a program (a sequence of primitive operations) that would lead to the tutor's answer. Once a plausible program is found, the learner can store this new utterance-program pairing as a candidate construction.

Over time, through an inductive process, the learner generalises across observed utterances and reconstructed meanings to build more abstract schemata (i.e. *pattern finding*). For example, if the learner has previously encountered and understood the question *"How many spheres are there?"* and then observes *"How many cubes are there?"*, it can induce a pattern. As shown in Figure 2, one possible generalisation could yield a construction that includes a slot, e.g. HOW-MANY-?X-ARE-THERE?-CXN, and another that can fill that slot, e.g. CUBES-CXN. A construction can thus be partially schematic: containing both fixed elements and variable slots. Slots are the parts that remain open and available to be filled by other constructions. Constructions may contain more than one slot, and slots can also occur adjacently. In the remainder of this text, we refer to partially schematic constructions with one or more slots as *item-based* constructions, while fully substantive constructions are referred to as *lexical* constructions.

As the construction inventory grows, the learner becomes able to interpret parts of novel utterances. The learner can then use this partial analysis as a starting point to more efficiently search for the remaining operations needed to construct a full program that leads to the answer. In total, seven generalisation operators are introduced by Nevens et al. (2022) and Doumen et al. (2024).

A critical component of the approach is the *categorial network* which organises the learner's acquired knowledge of which constructions can fill in slots of other constructions (Van Eecke, 2018). As seen in Figure 2, the *how-many-?x-are-there(?x)* category is linked to three filler categories (*spheres*, *cylinders*, *cubes*) that can fill the *?x* slot. The categorial network thus stores slot-filler relations observed during interactions and dynamically expands as new combinations are encountered. This mechanism supports an important generalisation: even when the learner has never seen a particular combination of constructions, it can still interpret the utterance if the individual components are known. For example, the learner might already know a construction WHAT-IS-THE-?X-MADE-OF?-CXN and another SPHERE-CXN, but never observed the specific combination *"What is the sphere made of?"*. In such cases, the categorial network allows the learner to combine known constructions by creating a new link between these categories, without needing to create a new construction.

Together, intention reading, pattern finding and the categorial network form the core mechanisms through which the learner agent acquires a flexible and compositional grammar. Through this grammar, the agent can solve the task of interpreting and answering the questions.

3 Acquiring a Construction Grammar from Sensory Experience

To demonstrate how a computational construction grammar spanning multiple levels of abstraction can be acquired directly from sensory experience, we integrate the experiments discussed in Sections 2.1 and 2.2. In our integrated methodology, a learner agent first acquires grounded lexical constructions through a reference-based game, before using these constructions as building blocks in a question-answering game.

To achieve this integration, the symbolic scene representations used in the question-answering game must be replaced by a continuous environment. As discussed in Section 2.2, the original experiment assumes symbolic input in the form of structured representations. This allows the learner to reason directly over discrete, high-level structures using its primitive operators, bypassing the challenge of perceptual grounding. In the continuous setting, the primitive operators must be adapted to reason over low-level structures which is not straightforward. We highlight three key changes.

Similarity between concepts and objects A crucial cognitive operation in the experiment is the FILTER primitive, which takes a set of objects as input and returns a filtered set containing only those objects for which a given concept applies. In the original symbolic setting, filtering objects by a concept relied on symbolic matching. In our continuous setup, we adapt the FILTER primitive to work

with raw perceptual features. Rather than checking whether an object has a given symbolic feature, the learner now computes a similarity score between the grounded concept and each object in the input set. This similarity is calculated using the algorithm introduced in Nevens et al. (2020). It estimates the likelihood that an object's sensory features were generated by the concept's distribution. Any object whose similarity exceeds a threshold γ is included in the filtered set.

Deriving category hierarchies The original question-answering experiments operate under a critical assumption: agents must also already have a prespecified *category hierarchy* (Rosch et al., 1976) to perform certain basic cognitive operations, such as querying on a category (e.g. COLOUR). The learner is assumed to already understand, for example, that SIZE constitutes a superordinate category with mutually exclusive values like SMALL and LARGE. This assumption provides a scaffold that simplifies the problem, but it does raise questions about how such hierarchies can be acquired.

It has been hypothesised that a categorial network, capturing slot-filler relationships, contains the information needed to derive these category hierarchies (Van Eecke, 2018; Steels et al., 2022; Nevens et al., 2022; Doumen et al., 2024). Simply put, categories that behave similarly across constructions may belong to the same domain. To operationalise this hypothesis, we identify potential semantic fields: groups of categories that likely belong to the same domain. This is achieved by clustering categories based on their *constructional behaviour* captured by the categorial network. Concretely, we compute the vertex cosine similarity (Salton and McGill, 1983) between all pairs of categories (i.e. fillers) in the categorial network. This yields a fully connected graph where each node corresponds to a category and each edge is weighted by the similarity score. Categories that frequently fill the same slots will have many shared connections, and thus a higher vertex cosine similarity. To identify meaningful clusters, we apply a threshold τ and retain only edges with high similarity scores. This pruning step breaks the network into connected components that represent potential semantic fields, such as size, colour or shape.

Generalisation operators As discussed in Section 2.2, the original question-answering experiment uses seven generalisation operators. These

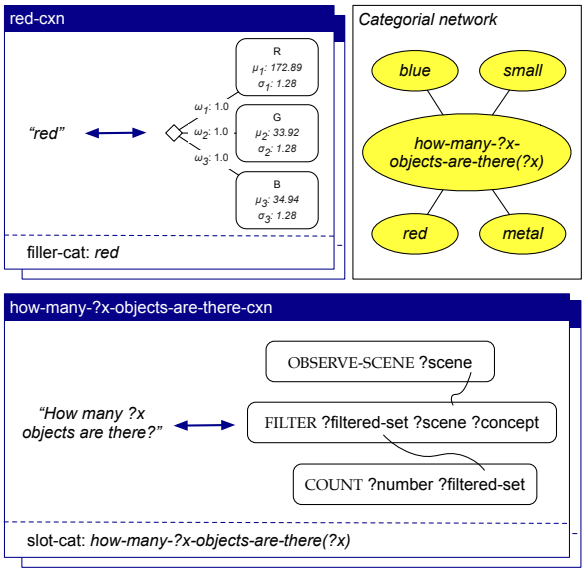


Figure 3: Example of two constructions acquired by the learner during the question-answering game in the continuous setting. A lexical RED-CXN, an item-based HOW-MANY-?X-OBJECTS-ARE-THERE-CXN and part of the categorial network, capturing the slot-filler relation through the categories in the constructions, are shown.

included generalisations over holistic mappings between linguistic forms and reconstructed meanings. These types of generalisations could lead to item-based and lexical constructions. In our experiment, all required lexical constructions are acquired before the question-answering game begins. As a result, operators that generalise over holistic mappings that yield lexical constructions are no longer needed. Due to our two-phased approach, only three generalisation operators are used (i.e. *add-categorial-link*, *lexical→item-based* and *nothing→holophrase*).

4 Experimental Validation

We validate our methodology experimentally. The experiment is structured in two phases. Initially, the learner acquires concepts through a reference-based game using the methodology discussed in Section 2.1. After this phase, the learner has acquired a set of grounded lexical constructions that are mappings between linguistic forms and perceptually grounded concept representations. In the second phase, the learner participates in a question-answering game using the adaptations discussed in Section 3. Concretely, the learner agent further expands its construction inventory with item-based constructions, in which the previously acquired

grounded lexical constructions serve as fillers. Figure 3 captures this idea: the construction inventory of the agent consists of both grounded lexical constructions as well as item-based constructions linked through the categorial network. In contrast to Figure 2, the meaning of the lexical construction is now represented by a grounded concept. Importantly, although the two phases are structured sequentially, learning is not confined to each phase: during the second phase, the concept representations in the grounded lexical constructions continue to be refined through new observations.

Data The experiment uses the CLEVR dataset (Johnson et al., 2017). This dataset contains questions about images containing three to ten geometric objects. Each object is described by a combination of attributes: one of three shapes (SPHERE, CUBE or CYLINDER), one of eight colours (GREY, BLUE, BROWN, YELLOW, RED, GREEN, PURPLE or CYAN), one of two material types (METAL or RUBBER) and one of two sizes (SMALL or LARGE).

Following Nevens (2022) and Doumen et al. (2024), we use a subset of the CLEVR scenes and questions. To create the continuous environment, we extract features for each object in the scenes from the dataset following the data processing steps discussed in Nevens et al. (2020, p. 7). We use 14,000 of the 15,000 scenes across both *Phase 1* and *2* and hold out the remaining 1,000 scenes for evaluation. This allows us to assess how well the methodology generalises to previously unseen scenes after *Phase 2*. Only questions involving (i) counting, (ii) checking for existence and (iii) querying for a certain attribute are retained. To obtain this subset, we removed questions related to comparison, spatial relations and logical operations. As explained in Doumen et al. (2024), this choice is motivated by the complexity of these operations which is far removed from the complexity of the questions that children encounter in the beginning of the language acquisition process. Lastly, in the CLEVR dataset, synonyms are used to describe the exact same concepts (e.g. sphere and ball). We remove these questions, following the principle of no synonymy (Goldberg, 1995, p. 67). Thus, in *Phase 2* of the experiment 1,935 unique questions can be posed about 14,000 different scenes.

Experimental setup The experimental setup of *Phase 1* follows Nevens et al. (2020). *Phase 2*, due to its increased complexity, is further broken down into three successive steps to facilitate learning. First, the tutor poses counting and existence check questions (respectively named *Phase 2A* and *Phase 2B*). This allows the learner to observe many slot-filler relationships and gradually build up its categorial network. After this, the tutor moves onto questions related to querying attributes of objects (*Phase 2C*), which requires reasoning over category hierarchies. These hierarchies are created based on the categorial network that was built up during the previous phases using the category clustering method described in Section 3. The thresholds γ and τ are hyperparameters and are set empirically to respectively $\gamma = 0.8$ and $\tau = 0.7$. In total the experiment consists of 20,000 interactions. *Phase 1* consists of 5,000 interactions, while *Phase 2* consists of 5,000 interactions for each of the three parts: *Phase 2A*, *2B* and *2C*. All reported results are averaged over 10 independent runs. All runs were conducted on a 12-core CPU paired with 16GB of RAM, with each run completed in ± 0.5 hours.

Learning dynamics The learning dynamics of the experiment are shown in Figure 4. For each phase, we keep track of the average communicative success over time. For the reference-based game, an interaction is successful if the learner points to the tutor's intended referent. For the question-answering game, there is success when the learner utters the expected answer. In both cases, a success of 100% means that learner understands the tutor perfectly.

As seen in Figure 4, at the end of *Phase 1*, an inventory of 15 grounded lexical constructions is acquired.[1] In the beginning of *Phase 2A*, when the learner encounters questions related to counting, success drops down, but quickly rises again when the learner successfully acquires item-based constructions that are needed to answer the questions. By the end of this phase, on average, 20 item-based constructions and 4 holophrase constructions are acquired and the necessary links between the slots of the item-based constructions and the slot-filler relations are learned and added to the categorial network. Similar dynamics are observed when the existence and query questions are introduced (*Phases 2B and 2C*). First, the success drops down, but the

[1] Note that the number of lexical constructions jumps from 15 to 18 between *Phase 1* and *Phase 2*. This increase is due to the creation of three plural equivalents for the singular 'shape'-constructions. In *Phase 1*, the tutor only refers to singular concepts, but later in the experiment, the plural versions are required.

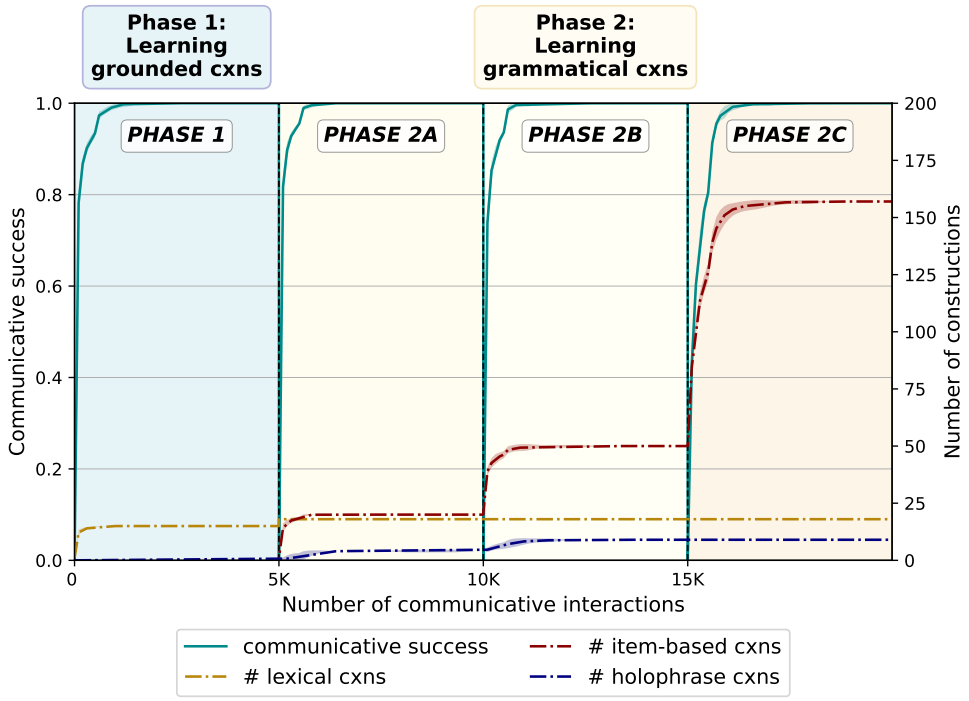


Figure 4: Learning dynamics of the experiment. The blue line denotes the degree of average communicative success over the past 500 interactions. At the start of each phase the average window is reset. The other dashed lines (yellow, red and blue) respectively denote the number of lexical, item-based and holophrase constructions that were acquired over time. At the beginning of each phase, communicative success drops down, but quickly recovers as constructions are acquired to resolve communicative impasses. Results are averaged over 10 independent runs.

agent quickly acquires the necessary constructions and communicative success is reached again after a couple of hundred interactions. The linguistic inventory size expands to $\pm$ 50 item-based and 9 holophrase constructions at the end of *Phase 2B* and reaches a number of 157 item-based constructions at the end of the experiment, leading to a total of $\pm$ 184 constructions.[2] Finally, we evaluate the acquired construction grammar on a held-out set of 1,000 unseen scenes in ten independent runs. During this phase, the learner's linguistic system is frozen and cannot be updated. We perform 5,000 additional interactions on this evaluation set. The learner correctly interprets and answers the tutor's question posed in 99.65% of interactions, averaged over 10 independent runs. Analysis of the rare failure cases reveals that errors are primarily due to grounding issues, where slight out-of-distribution observations relative to the learned concept representations lead down the line to incorrect answers.

Formation of a category hierarchy The categorial network captures the slot-filler relations of the constructions. These relations are built up during the experiment and form the basis for the formation of category hierarchies, which are used in the last phase of the experiment.

Figure 5 shows the expansion of the learner's categorial network during the different phases of the experiment. For visual purposes, we zoom in on categories related to nine grounded lexical constructions and three item-based constructions. After *Phase 1*, the network consists only of disconnected categories for grounded lexical constructions. During *Phase 2A* categories start to cluster. We observe that categories that relate to the shape of objects act as fillers in similar slots (e.g. they fill the *?y* slot in the HOW-MANY-?X-?Y-ARE-THERE-CXN) and are thus possibly more related to each other than, for example, the 'material', 'colour' and 'size' categories which fill the *?x* slot in the same construction (see Figure 5). During *Phase 2B*, the categorial network expands. Now, we clearly ob-

[2]Note that there is no typical 'overshoot' pattern for the number of constructions. In the reference-based game, this is due to the lack of ambiguity regarding form about which forms map to which meaning. The learner directly acquires a construction with an initial concept representation that is gradually refined. In the question-answering game, we observe that the agent likewise acquires the optimal meaning representation from the start. This contrasts with the original experiment, where many suboptimal lexical mappings were first acquired. Our two-phased approach prevents lexical suboptimal hypotheses.

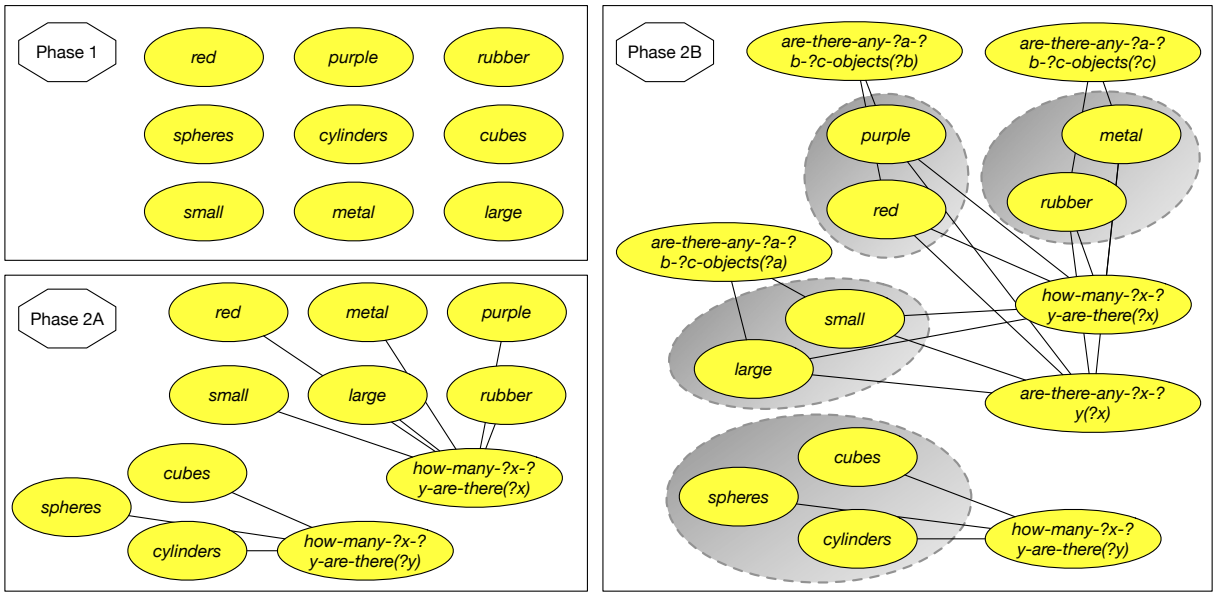


Figure 5: Expansion of the learner's categorial network over the course of the experiment. *Phase 1* shows the network after the grounded lexical construction learning phase, with no links between categories yet. In *Phase 2A* initial clusters of categories begin to form. In *Phase 2B*, shaded in grey, four semantically relevant clusters emerge (size, colour, material, shape). Only a subset of the categorial network is shown for illustrative purposes.

serve that categories cluster together into meaningful hierarchies over concepts, indicated by the grey shaded regions in Figure 5. This cluster formation is used to build the hierarchy needed in *Phase 2C* of the experiment in which the agent needs to query attributes of certain objects. Applying the methodology described in Section 3 results in 5 clusters: the shapes (singular and plural), the colours, the materials and the sizes. Our results demonstrate that a useful category hierarchy can emerge based on the constructional behaviour captured by the categorial network of an agent.

5 Related Work

Many computational models for grounded language acquisition have been developed in different fields, including cognitive linguistics, AI and robotics. This paper studies this grand challenge from a constructionist perspective. Therefore, in what follows, we outline the different strands of work that take this perspective. Following a recent survey by Doumen et al. (2025), we thus focus on constructionist models that incrementally acquire productive form-meaning mappings that extend beyond lexical items. Doumen et al. (2025) distinguish approaches based on how much semantic supervision is provided. In a first set of models, training examples pair an utterance with its gold semantic annotation (e.g. Alishahi and Stevenson, 2008; Beuls et al., 2010; Chang, 2008; Doumen et al., 2024; Gerasymova and Spranger, 2010, 2012). Other models reduce this supervision by presenting multiple candidate gold semantic annotations, introducing referential uncertainty (Abend et al., 2017; Beekhuizen and Bod, 2014; Beekhuizen, 2015; Chen and Mooney, 2008; Dominey, 2005a,b; Dominey and Boucher, 2005; Garcia Casademont and Steels, 2015, 2016; Gaspers et al., 2011; Gaspers and Cimiano, 2012, 2014; Gaspers et al., 2017; Kwiatkowski et al., 2010, 2011, 2012; Steels, 2004). A third set of models focus on learning in situated interactions without gold semantic annotations altogether. In these works, a combination of a predefined lexicon, categories or ontology is assumed (Artzi and Zettlemoyer, 2013; Nevens et al., 2022; Spranger, 2015; Spranger and Steels, 2015; Spranger, 2017). Finally, De Vos et al. (2024) present a grammar coupled with concepts grounded in a way similar to Section 2.1. Notably, their approach is applied to the same visual question answering task considered in this paper. However, whereas they manually designed a grammar tailored to the task, our focus lies on the acquisition of a grammar across different levels of abstraction. This makes the problem significantly more challenging and motivated our use of a subset of the dataset (see Section 4). As such, direct comparison is not straightforward. De Vos

et al. (2024) report an accuracy of 96% on the full dataset, while we achieve near-perfect success on the subset.

A growing related strand of research examines to what extent large language models (LLMs) capture constructional knowledge. These probing studies indicate that state-of-the-art LLMs can capture substantive constructions reasonably well, but have more difficulty with more schematic patterns (see e.g. Weissweiler et al. (2022); Bonial and Tayyar Madabushi (2024); Zhou et al. (2024); Rozner et al. (2025)). These findings provide valuable insights into the strengths and limitations of current models. Our objective, rather, is to present a mechanistic model in which constructions at varying levels of schematicity emerge incrementally through situated communicative interactions, rather than via optimisation for next-token prediction over large-scale corpora.

6 Discussion and Conclusion

This paper has presented a computational model in which a construction grammar is acquired directly from sensory experience, capturing constructions at varying levels of abstraction. We have integrated two previously separate but related experiments operationalised in the language game paradigm, guided by the hypothesis that constructions at different levels can be acquired through the same underlying cognitive mechanisms. While Beuls and Van Eecke (2024) formulated this idea at a conceptual level, we offer a concrete operationalisation. In our approach, constructions are acquired through repeated situated communicative interactions between a tutor and a learner agent. Across these interactions, the learner identifies regularities (whether these are associations between sensory feature values and linguistic forms or correspondences between syntactic patterns and semantic operations) and uses those regularities to incrementally refine its linguistic system. To enable this integration, we have introduced a component that induces a category hierarchy from the slot-filler relations of the acquired constructions, thereby replacing a major scaffold of the earlier model by Nevens et al. (2022), which assumed access to a predefined hierarchy. In this setting, the component derives category hierarchies that are one layer deep, although future work could investigate extensions to multi-level hierarchies.

The methodology has been validated through an experiment in the synthetic CLEVR environment. The experiment has demonstrated that lexical constructions that were grounded in the sensors of the agent and were acquired in referential games can serve as building blocks for abstract grammatical constructions in a subsequent question-answering game. In this paper, we focused on the acquisition of lexical and item-based constructions. These results provide computational support for a core assumption in construction grammar, showing how both purely substantive and more abstract constructions can emerge from repeated situated communicative interactions. However, further work is needed to investigate the acquisition of constructions at all levels of abstraction in more complex environments.

Acknowledgements

We would like to thank Liesbet De Vos, Jamie Wright, Arno Temmerman and the anonymous reviewers for their valuable comments on earlier versions of the paper. The research reported on in this paper was funded by the F.R.S.-FNRS-FWO WEAVE project HERMES I under grant numbers T002724F (F.R.S.-FNRS) and G0AGU24N (FWO), the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme and the AI Flagship project ARIAC by DigitalWallonia4.ai.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science*, 32(5):789–834.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

Barend Beekhuizen. 2015. *Constructions Emerging*. Ph.D. thesis, Universiteit Leiden.

Barend Beekhuizen and Rens Bod. 2014. Automating construction work: Data-oriented parsing and onstructivist accounts of language acquisition. In Ronny Boogaart, Timothy Colleman, and Gijsbert Rutten, editors, *Extending the Scope of Construction Grammar*, pages 47–74. De Gruyter Mouton, Berlin, Germany.

Heike Behrens. 2009. Usage-based and emergentist approaches to language acquisition. *Linguistics*, 47(2):383–411.

Katrien Beuls, Kateryna Gerasymova, and Remi van Trijp. 2010. Situated learning through the use of language games. In *Proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands (BeNeLearn)*, pages 1–6.

Katrien Beuls and Sebastian Höfer. 2011. Simulating the emergence of grammatical agreement in multi-agent language games. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pages 61–66, Washington, D.C., USA. AAAI Press.

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. What about machines? *Computational Linguistics*, 50(4):1277–1311.

Katrien Beuls and Paul Van Eecke. 2025. Construction grammar and artificial intelligence. In Mirjam Fried and Kiki Nikiforidou, editors, *The Cambridge Handbook of Construction Grammar*, pages 543–571. Cambridge University Press, Cambridge, United Kingdom.

Claire Bonial and Harish Tayyar Madabushi. 2024. Constructing understanding: on the constructional information encoded in large language models. *Language Resources and Evaluation*.

Jérôme Botoko Ekila, Jens Nevens, Lara Verheyen, Katrien Beuls, and Paul Van Eecke. 2024. Decentralised emergence of robust and adaptive linguistic conventions in populations of autonomous agents grounded in continuous worlds. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-Agent Systems*, pages 2168–2170, Richland, SC, USA. IFAAMAS.

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

Nancy Chang. 2008. *Constructing grammar: A computational model of the emergence of early constructions*. Ph.D. thesis, University of California, Berkeley, Berkeley, CA, USA.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine learning*, pages 128–135.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, Oxford, United Kingdom.

Joachim De Beule and Benjamin K. Bergen. 2006. On the emergence of compositionality. In *The Evolution of Language. Proceedings of the 6th International Conference (EVOLANG6)*, pages 35–42, Singapore, Singapore. World Scientific.

Liesbet De Vos, Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2024. Construction grammar and procedural semantics for human-interpretable grounded language processing. *Linguistics Vanguard*, 10(2):565–574.

Holger Diessel. 2017. Usage-based linguistics. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Oxford, United Kingdom.

Peter Ford Dominey. 2005a. Emergence of grammatical constructions: Evidence from simulation and grounded agent experiments. *Connection Science*, 17(3-4):289–306.

Peter Ford Dominey. 2005b. From sensorimotor sequence to grammatical construction: Evidence from simulation and neurophysiology. *Adaptive Behavior*, 13(4):347–361.

Peter Ford Dominey and Jean-David Boucher. 2005. Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, 167(1):31–61.

Jonas Doumen, Katrien Beuls, and Paul Van Eecke. 2024. Modelling constructivist language acquisition through syntactico-semantic pattern finding. *Royal Society Open Science*, 11(7):231998.

Jonas Doumen, Veronica J. Schmalz, Katrien Beuls, and Paul Van Eecke. 2025. The computational learning of construction grammars: State of the art and prospective roadmap. *Constructions and Frames*, 17(1):141–174.

Ewa Dąbrowska and Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, 16(3):437–474.

Charles J. Fillmore. 1988. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 35–55.

Emília Garcia Casademont and Luc Steels. 2015. Usage-based grammar learning as insight problem solving. In *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science*, pages 258–263.

Emília Garcia Casademont and Luc Steels. 2016. Insight grammar learning. *Journal of Cognitive Science*, 17(1):27–62.

Judith Gaspers and Philipp Cimiano. 2012. A usage-based model for the online induction of constructions from phoneme sequences. In *Proceedings of the 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–6. IEEE.

Judith Gaspers and Philipp Cimiano. 2014. A computational model for the item-based induction of construction networks. *Cognitive Science*, 38(3):439–488.

Judith Gaspers, Philipp Cimiano, Sascha S. Griffiths, and Britta Wrede. 2011. An unsupervised algorithm for the induction of constructions. In *Proceedings of the 2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–6. IEEE.

Judith Gaspers, Philipp Cimiano, Katharina Rohlfing, and Britta Wrede. 2017. Constructing a language from scratch: Combining bottom–up and top–down learning processes in a computational model of language acquisition. *IEEE Transactions on Cognitive and Developmental Systems*, 9(2):183–196.

Kateryna Gerasymova and Michael Spranger. 2010. Acquisition of grammar in autonomous artificial systems. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI-2010)*, pages 923–928. IOS Press.

Kateryna Gerasymova and Michael Spranger. 2012. An experiment in temporal language learning. In Luc Steels and Manfred Hild, editors, *Language Grounding in Robots*, pages 237–254. Springer, New York, NY, USA.

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

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

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.

Frédéric Kaplan, Luc Steels, and Angus McIntyre. 1998. An architecture for evolving robust shared communication systems in noisy environments. Technical report, Sony CSL Paris.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 234–244. Association for Computational Linguistics.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1223–1233. Association for Computational Linguistics.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical generalization in CCG grammar induction for semantic parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1512–1523. Association for Computational Linguistics.

George Lakoff. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford University Press, Stanford, CA, USA.

Elena Lieven. 2014. First language learning from a usage-based approach. In Thomas Herbst, Hans-Jörg Schmid, and Susen Faulhaber, editors, *Constructions Collocations Patterns*, pages 9–32. De Gruyter Mouton, Berlin, Germany.

Martin Loetzsch. 2015. *Lexicon formation in autonomous robots*. Ph.D. thesis, Humboldt-Universität zu Berlin, Berlin, Germany.

Jens Nevens. 2022. *Representing and learning linguistic structures on the conceptual, morphosyntactic, and semantic level*. Ph.D. thesis, Vrije Universiteit Brussel, Brussels: VUB Press.

Jens Nevens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2022. Language acquisition through intention reading and pattern finding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–25, Gyeongju, Republic of Korea.

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2020. From continuous observations to symbolic concepts: A discrimination-based strategy for grounded concept learning. *Frontiers in Robotics and AI*, 7(84).

Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025. Constructions are revealed in word distributions. *arXiv preprint arXiv:2503.06048*.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw Hill.

Michael Spranger. 2015. Incremental grounded language learning in robot-robot interactions: Examples from spatial language. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 196–201. IEEE.

Michael Spranger. 2017. Usage-based grounded construction learning: A computational model. In *The 2017 AAAI Spring Symposium Series*, pages 245–250, Washington, D.C., USA. AAAI Press.

Michael Spranger and Luc Steels. 2015. Co-acquisition of syntax and semantics: an investigation in spatial language. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1909–1915, Washington, D.C., USA. AAAI Press.

Luc Steels. 1995. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332.

Luc Steels. 1999. *The Talking Heads experiment: Volume I. Words and Meanings*. Best of Publishing, Brussels, Belgium.

Luc Steels. 2004. Constructivist development of grounded construction grammar. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 9–16. Association for Computational Linguistics.

Luc Steels and Emília Garcia Casademont. 2015. Ambiguity and the origins of syntax. *The Linguistic Review*, 32(1):37–60.

Luc Steels, Paul Van Eecke, and Katrien Beuls. 2022. Usage-based learning of grammatical categories. *arXiv preprint arXiv:2204.10201*.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Harvard, MA, USA.

Paul Van Eecke. 2018. *Generalisation and specialisation operators for computational construction grammar and their application in evolutionary linguistics Research*. Ph.D. thesis, Vrije Universiteit Brussel, Brussels: VUB Press.

Remi van Trijp. 2008. The emergence of semantic roles in fluid construction grammar. In *The Evolution of Language*, pages 346–353. World Scientific.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882. Association for Computational Linguistics.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

William A. Woods. 1968. Procedural semantics for a question-answering machine. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, pages 457–471, New York, NY, USA.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions are so difficult that even large language models get them right for the wrong reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811. Association for Computational Linguistics.

Performance and competence intertwined: A computational model of the Null Subject stage in English-speaking children

Soumik Dey
The Graduate Center
The City University of New York
sdey@gradcenter.cuny.edu

William Gregory Sakas
The Graduate Center
The City University of New York
wsakas@hunter.cuny.edu

Abstract

The empirically established null subject (NS) stage, lasting until about 4 years of age, involves frequent omission of subjects by children. Orfitelli and Hyams (2012) observe that young English speakers often confuse imperative NS utterances with declarative ones due to performance influences, promoting a temporary null subject grammar. We propose a new computational parameter to measure this misinterpretation and incorporate it into a simulated model of obligatory subject grammar learning. Using a modified version of the Variational Learner (Yang, 2012) which works for superset-subset languages, our simulations support Orfitelli and Hyams' hypothesis. More generally, this study outlines a framework for integrating computational models in the study of grammatical acquisition alongside other key developmental factors.

1 Introduction

The Null Subject (*NS*) stage is a well-researched phenomenon in child language acquisition, characterized by young children sometimes forming declarative sentences without subjects. This is expected in children exposed to null subject languages but contentious in obligatory subject language environments. The NS stage challenges the Subset Principle (Gold, 1967; Berwick, 1985; Manzini and Wexler, 1987; Valian, 1990; Déprez and Pierce, 1993; Fodor and Sakas, 2005) — children learning obligatory subject languages exhibit NS-like sentences (a superset language), which gradually shift to non-NS (subset language) with time. This phenomenon puzzles learning theorists. Explanations vary, with some attributing the NS stage to differences between children's internal grammar and adult target grammars (Yang, 2012; Orfitelli and Hyams, 2008; Valian, 1990), while others cite extrasyntactic factors like memory and processing constraints (Bloom, 1970, 1990; Valian, 1991; Wang et al., 1992). Rizzi (2005a,b) connects a performance account of a limited production system with its consequence of the varying grammatical competence we see in children. In this paper, we model the grammatical theory of the NS stage in children using a developmental parameter and the Variational Learner (VL)(Yang, 2012), a well-known computational model of language acquisition. More generally, this study outlines a framework for integrating computational models in the study of language acquisition alongside other key developmental factors.

2 Background

2.1 Orfitelli & Hyams (2012) Experiment 2

The two distinct theories of performance and grammatical competence present distinct explanations for children's comprehension of subject-lacking sentences (*NS sentences*, such as imperatives in English). Grammatical theories propose that young English speakers view NS sentences akin to grammatically correct declaratives, similar to adults in null subject languages. Conversely, performance theories attribute omissions to production limitations, suggesting children interpret NS sentences as adults do in obligatory subject languages, limiting English-speaking children's interpretations to imperatives or diary forms. To explore this, Orfitelli and Hyams (2012, Experiment 2) used a truth-value judgment (TVJ) experiment (Crain and McKee, 1985; Crain and Fodor, 1993). In Experiment 2, a child watched a narrative, then listened to a puppet's (Mr. Bear) comment, and judged the comments' accuracy relative to the story. The child corrected Mr. Bear by indicating the correctness of his statements, with explanations. Thirty children from Los Angeles daycare centers were involved, divided into three age groups (2;6-2;11, 3;0-3;5, and 3;6-3;11) to represent early, middle, and late NS stages. Details on age range and distribution

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 96–108

Group	Age Range	Mean Age	N
2;6-2;11	2.54-2.96	2.73	10
3;0-3;5	3.12-3.48	3.3	10
3;6-3;11	3.64-3.98	3.82	10
Total	2.54-3.98	3.28	30

Table 1: Orfitelli and Hyams (2012)[Experiment 2] participant details.

are in Table 1, adapted from Orfitelli and Hyams (2012, Table 6).

The children underwent assessment on 24 grammatical items (sentences), equally split between correct and incorrect true/false responses. There were 8 NS condition sentences, while the remaining 16 items were evenly divided among the remaining four conditions. Orfitelli & Hyams (O&H) classified the children's responses to NS condition sentences into three categories based on interpretation:

- **Consistently imperative**: 7-8 out of 8 NS sentences interpreted as imperative.
- **Both interpretations allowed**: 2-6 out of 8 NS sentences interpreted as imperative.
- **Consistently declarative**: 0-1 out of 8 NS sentences interpreted as imperative.

	2;6-2;11	**3;0-3;5**	**3;6-3;11**
Imperative (7-8 imp)	0% (0)	40% (4)	80% (8)
Both (2-6 imp.)	80% (8)	60% (6)	20% (2)
Declarative (0-1 imp.)	20% (2)	0% (0)	0% (0)

Table 2: Individual performance on the NS condition sentences in Orfitelli and Hyams (2012)[Table 8].

Additional details regarding the performance of children on the task are illustrated in Figure 1. The performance on the NS items varied with age as O&H reported that the youngest group assigned an imperative (adult) interpretation to NS items 40% of the time on average, while the middle age group assigned an imperative interpretation 64% of the time. While O&H do not report an average number for the performance on NS items for the oldest age

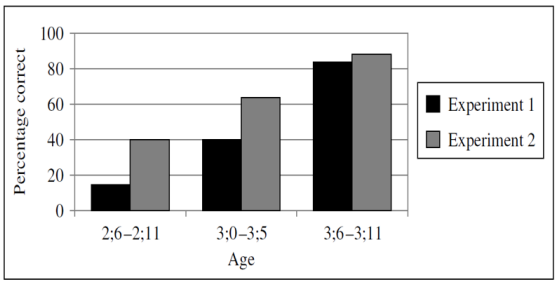


Figure 1: Performance on NS condition sentences from Orfitelli and Hyams (2012)[Figure 5].

group, from Figure 1 we can estimate the average performance to be close to 90%. For concreteness, we adopt 90% for this age group from this point on.

The fact that in O&H's study children comprehend NS sentences differently than an adult, reinforces the grammatical account of the NS stage. However, O&H also argue for performance limitations which create the illocutionary force ambiguity associated with the imperative NS sentences resulting in the NS stage.

2.2 Language Acquisition in P&P Framework

The study of language acquisition presents an extraordinary challenge for scientific inquiry. It requires that a child, over a remarkably short period, must develop a grasp of a grammar system capable of producing and interpreting a set of utterances comparable to those produced by adults within their linguistic surroundings.[1] The child's cognitive mechanisms for language learning achieve this despite having limited or no exposure to sentence-level linguistic phenomena, and without the capacity to perceive intrinsic properties of the latent structures that generate the surface forms of utterances (see Fodor 1998 for reference). This restricted interaction with sufficient surface forms forms the core of the argument known as the poverty of the stimulus (Chomsky, 1981, 1955, 1965, 1986).

This argument has been instrumental in advocating for an intrinsic language faculty that imparts universal structural principles (such as the notion that all languages possess subjects) and parameters, which dictate language-specific structural traits (e.g., whether SpecIP is initial or final) that are adjusted during the language acquisition process.,The framework of principles and parameters

[1] Or nearly identical, encompassing microvariations within linguistic communities.

(P&P), as introduced by (Chomsky, 1981), was designed to streamline the language learning process by:

- Limiting the potential scope of grammatical possibilities, transitioning linguistic theory from a potentially limitless universe of human grammars to a explicitly finite set
- Simplifying complex structural phenomena into parameter values, which vary between languages. This framework comprises a set of foundational principles that "sharply restrict the class of attainable grammars and narrowly constrain their form, but with parameter [values] that have to be fixed by experience" (Chomsky, 1981).

Essentially, the child is inherently equipped with these principles, while parameter values are influenced by the linguistic input they encounter in their environment.[2] We align with Fodor's interpretation of parameter values within the P&P model (Fodor and Sakas 2005): Universal Grammar (UG) endows parameters with two possible, albeit mutually exclusive, structural "treelets" – elements of grammatical architecture – that serve as tools for both linguists and children acquiring language to distinguish between different human languages. Subsequently, (Howitt et al., 2021) suggest that parameter values should be seen *not* as simple binary choices between parametric treelets, but rather as points within a gradient spectrum between these discrete choices, viewing parameter values as dynamically adjustable along a continuum.

2.3 Variational Learner

Yang (2002a, 34) argues for the necessity of learners to perform well in domains without unambiguous inputs (see Clark 1992; Clark and Roberts 1993 who argue against the general existence of unambiguous evidence). He proposes a parameter setting reward-based algorithm that converges to a target grammar despite the presence of ambiguous evidence (Straus, 2008). His *Variational Learning* (*VL*) model posits that a child accesses multiple grammars, competing throughout learning. When encountering a sentence, the child uses her current grammar hypothesis for parsing. Success results in rewards; failure incurs penalties. Competing grammars vie to become the next hypothesis, with the most rewarded becoming the adult grammar. In VL, a *learning rate* $\hat{R}$ dictates grammar rewards or penalties. Each grammar G_i is linked to a probability P_i, indicating past rewards and penalties. At time t, this probability P_i depends on linguistic exposure E_t and grammar performance. Implementing variational learning with Principles and Parameters involves managing 2^n probabilities for grammars in an *n*-parameter space, exceeding a billion in a 30-parameter P&P domain. Yang suggests maintaining one weight (w_i) per parameter (p_i). Like non-parametric VL, parameters are adjusted based on parsing outcomes, modifying weight (w_i) accordingly. Each p_i is binary, with value (p_i^v) of 0 or 1. Grammar probabilities form a weight vector (W) of size n, where w_i aligns with parameter p_i. Weights encode cumulative parametric reward and penalty results at time t after E_t. In P&P VL, Yang (2002a) details two weight update methods following a sentence parse (s_t) at time t. Weights vector $W = [w_1, w_2 ... w_n]$ is adjusted, rewarding successful parsing by $G_{curr} = [p_1^v, p_2^v ... p_n^v]$ and penalizing failures. Updated weights then define new G_{curr}. Yang (2012) describes a *reward-only* VL where unsuccessful parsing leaves weights unaltered. Following Sakas et al. (2017), we adopt and modify principles and parameters reward-only VL for simulations.

The reward scheme of the reward-only VL follows the (L_{R-P}) scheme of Bush and Mosteller (1955). If a parameter value, p_i^v, in G_{curr} is 0 and w_i is to be rewarded, the weight is nudged towards 0 according to Equation (1):

$$w_i^{t+1} = w_i^t - \hat{R} \cdot (w_i^t) \tag{1}$$

If a parameter value, p_i^v, in G_{curr} is 1 and w_i is to be rewarded, the weight is nudged towards 1 according to Equation (2):

$$w_i^{t+1} = w_i^t + \hat{R} \cdot (1 - w_i^t) \tag{2}$$

Where w_i^t denotes weight w_i in the vector of weights W at time instance t. w_i^{t+1} is the weight after the update when encountering the input sentence s_t at time instance t.

Yang (2002b) hypothesized that a child is unable to distinguish between English grammar and its NS counterpart early on (imperfect learning), while in later stages of acquisition the corrective force of grammar competition sets the target parameter correctly.

[2] For the sake of simplicity, linguistic learnability typically presumes the language environment as monolingual.

3 Computational Modelling of NS utterance interpretation

The remainder of this paper presents a simulation study which models the work of Orfitelli and Hyams (2012). Specifically, we model the increase of a child's ability to interpret imperative sentences in an adult manner and observe the change in a simulated learner's (an *e-child*'s) competence over the course of language acquisition in an English-like abstract linguistic environment. We run simulation experiments employing a computational models of syntactic parameter setting: The Variational Learner (Yang, 2002a, 2012; Sakas et al., 2017) which we modify to incorporate (a version of) the Subset Principle. These experiments are run on the English-like language drawn from a large domain developed at the City University of New York (CUNY). The study presents a computational investigation of how performance factors might influence competence longitudinally.

3.1 The CUNY-CoLAG language domain

The CUNY-CoLAG domain is a database of word order patterns that children could be expected to encounter, together with all syntactic derivations of those patterns and the syntactic parameter values which generated each derivation. The multi-language domain is large, containing 3,072 artificial languages, 48,077 distinct word order patterns, and 93,768 distinct syntactic trees. Germane to this article, is CoLAG English (Sakas et al., 2017), most English-like language in the domain. A more thorough overview of the domain and how the multilingual derivations were generated can be found in Sakas (2003) and the most recent version of the CUNY CoLAG domain (hereafter, simply CoLAG) is comprehensively presented in Sakas and Fodor (2011, 2012).[3] [4]

3.2 Subset-superset languages and the Variational Learner

Yang's Variational Learner is highly regarded in terms of bringing statistical methods to the table together with generative grammar. However, the VL cannot distinguish between superset and subset grammars and cannot be prevented from converging on an incorrect superset hypothesis. However, a version of Yang's learner that *does* distinguish between superset and subset grammars and avoid convergence on an incorrect superset hypothesis can be envisioned: Whenever the learner encounters a sentence licensed by a current grammar hypothesis which generates a superset language, it checks if the sentence can be parsed by a subset hypothesis of the current grammar. If the sentence can be parsed by the subset grammar the learner picks the subset grammar choice, rather than the current (superset) grammar hypothesis for adjusting the weights (Yang, p.c.).

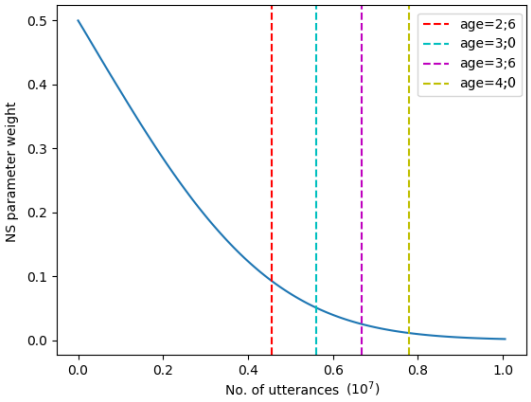


Figure 2: The SSVL with a conservative learning rate of $r = 1.24 \times 10^{-7}$. The NS parameter weight is plotted on the y-axis and the number of utterances on the x-axis. Additionally, 6 month intervals from age 2;6 to 4;0 as measured in number of utterances are marked.

We embrace this strategy, however, we found a need to augment it. The strategy focuses on acquiring a target subset grammar and is potentially detrimental, in the worst case fatally, when the VL is faced with a superset target grammar. Suppose an e-child employing the VL is trying to learn a superset target grammar. Every time the e-child hears an utterance that can be parsed by the subset grammar, the learner adjusts its weights in the direction of the subset grammar. Thus, convergence towards the superset is dependent on the order, and the ratio of sentences unambiguously licensed by the superset grammar to those licensed by the subset grammar.

To confirm our suspicions we ran this version of the Variational Learner with a 100 e-children acquiring CoLAG Null Subject English (NS-English), i.e., a language that has all the CoLAG English parameter settings except for Null Subject which has

[3] The domain is available for download at: `https://bit.ly/3nGdhPc`.

[4] While we acknowledge that CoLAG is an artificial domain, natural language domains like CHILDES(MacWhinney, 2000) has been explored with the VL in (Sakas et al., 2017). The CoLAG domain is used to prove a theoretical point and test the convergence pattern of model with a wide variety of distributions reminiscent to the study in (Howitt et al., 2021).

Algorithm 1 Superset-Subset Yang's Variational Learner reward only.

W is the array of weights G_{curr} is the current grammar, i.e. vector of parameter values
n is the number of parameters
for *each* w_i *in* W **do**
  $w_i \leftarrow 0.5$
end
for *each input sentence* s **do**
  pick $G_{curr} \leftarrow [p_1^v, ... , p_n^v]$ according to Algorithm 3 **if** G_{curr} *can parse* s **then**
    for w_i *in* W **do**
      if p_i *is a not superset-subset parameter or* p_i^v *is the subset value* **then**
        Adjust w_i conservatively towards p_i^v
      else // p_i^v `is the superset value`
        $G_{temp} \leftarrow [p_1^v, ..., 1 - p_i^v, ... p_n^v]$ // $1 - p_i^v$`=subset value`
        if G_{temp} *can parse* s **then**
          Adjust w_i conservatively towards $1 - p_i^v$;
        else
          Adjust w_i aggressively towards p_i^v;
        end
      end
    end
end

a value of 1 allowing null subjects in declaratives. All 100 e-children converged incorrectly on the subset value of the Null Subject parameter.[5]

We propose an adaptation of the VL which allows it to consistently converge to the correct parameter setting of a superset-subset parameter. The approach we adopt is to ensure that whenever the VL encounters a sentence that can be parsed only by the superset grammar, we reward it at a higher rate in comparison to the rate used for the subset value. The idea is to have two learning rates — a higher rate for rewarding the superset and a lower rate for rewarding the subset. Following Howitt et al. (2021), we will call them the "aggressive" (R) and "conservative" learning rates (r) respectively. To test this idea, we again ran simulations involving a 100 e-children acquiring CoLAG NS-English with learning rates of $R = 0.008$ and $r = 0.001$.[6] This learning strategy is successful — the *Superset-Subset Variational Learner* (*SSVL*) successfully converges on the target superset value for the Null Subject parameter for all e-children acquiring CoLAG NS-English, see Figure 2.

[5]Following (Sakas et al., 2017), the simulations were run on a uniform distribution of CoLAG English sentences with a learning rate of 0.001 and successful convergence was defined as the weights reaching within 0.02 threshold of the target parameter values.

[6]In line with Footnote 5, the aggressive rate of 0.008 was chosen through trial and error for the learner to converge with a conservative rate of 0.001.

Pseudocode for the SSVL is given in Algorithm 1. The initialization of the weights and the pick of G_{curr} is identical to Algorithm 3 (Yang's Reward-only VL). After every sentence, if the input sentence can be parsed by G_{curr}, the SSVL checks all p_i^v in G_{curr} for superset-subset values, if any. If p_i is not a superset-subset parameter or if p_i is a superset-subset parameter and p_i^v is the subset parameter value, w_i is rewarded conservatively towards p_i^v. Otherwise, p_i^v is a superset value. In that case, the SSVL checks if the current grammar with the superset value of p_i flipped (G_{temp} in Algorithm 1) to the subset value $1 - p_i^v$ can parse the current input sentence, if so, w_i is rewarded conservatively towards the subset value, Otherwise w_i is rewarded aggressively towards the superset value. As with the original reward-only VL, if the current input sentence can not be parsed by G_{curr} no weight updates occur.

The weights in W are rewarded as follows:

- **Reward aggressively**: Replace $\hat{R}$ by the aggressive rate R in Equation (1) or (2) and update w_i accordingly.
- **Reward conservatively**: Replace $\hat{R}$ by the conservative rate r in Equation (1) or (2) and update w_i accordingly.

The original Variational Learner follows the Naive Parameter Learning (NPL) model, which assumes that when the composite grammar successfully parses the incoming sentence, all parameter values are rewarded. However, as seen in our experiments involving CoLAG English and NS-English, for successful convergence on either the superset or subset grammar, the VL cannot not afford to be "naive". Specifically, it requires knowledge of which parameter values are in superset-subset relationship and exactly how to reward the relevant value.

3.3 A performance parameter: IARC

Orfitelli and Hyams (2012) based on their TVJ experiment, observe that there is a misinterpretation of illocutionary force in null subject sentences due to performance limitations in children and conjecture that adults and children have different grammars. This section outlines the modeling approach we adopt to capture this observation.

Table 2 presents data that show that children's ability to correctly interpret imperative illocutionary force changes over time. This change is almost

linear: Children between ages 2;6-2;11 show 40% adult interpretation of NS utterances on average, while children between ages 3;0-3;5 show 64% and 3;6-3;11 show 90%.

'We computationally model the interpretation of the illocutionary force of an e-child by introducing imperative NS sentences labeled with declarative illocutionary force into the e-child's linguistic environment. This "*noisy*" input to an e-child can be manipulated to mirror the data in Table 2 by decreasing the noise as the e-child matures. Employing this simulated performance factor, we map the pathway the NS parameter takes during the acquisition of CoLAG English. The question we are asking is — Assuming English children do indeed have a declarative interpretation of imperative NS sentences — how can we model the change in the Null Subject parameter, a parameter whose acquisition is affected by these NS sentences, to come to a conclusion regarding its target setting? And given the projected trajectory of this developmental change, what course would the trajectory of NS parameter acquisition take?

In learning CoLAG English, reliance must be placed on declarative utterances with subjects. Misunderstanding imperative NS forms as declaratives impairs learning, treating some NS utterances as noise and incorrectly shifting the parameter toward the null subject superset. In obligatory subject languages for adults, mature children's transition should reflect a shift from superset (optional subject) to subset (obligatory subject) grammars. Drawing on Orfitelli and Hyams (2012)'s TVJ experiment, we introduce the *Illocution Ambiguity Resolution Coefficient* (IARC) for measuring children's misinterpretations of imperatives. An IARC of 1 indicates perfect recognition of imperative NS as such, whereas an IARC of 0 signifies total misinterpretation as declaratives. An IARC of 0.2 suggests 20 out of 100 NS imperatives are correctly understood, with 80 misunderstood as declaratives. Our goal is to explore NS parameter acquisition in CoLAG English, considering these performance limitations.

3.4 Growth of IARC

In this section, we develop a framework for quantifying how the performance parameter IARC grows as a function of age, measured here by the cumulative utterances heard by a child at the end of age range i (U_i). Recall that $IARC$ is a probability measure and hence is bound within the values 0 and 1. As discussed in Section 2.1, the average values of IARC between the age ranges of 2;6-2;11, 3;0-3;5 and 3;6-3;11 exhibit almost linear growth (0.4 to 0.64 to 0.9). Thus, a natural way to model IARC would be as a bound function of U_i, $0 \leq IARC \leq 1$ with IARC linearly increasing with respect to U_i. One such approach is presented in Equation (3), where m is the slope, and c is the intercept of a linear function of IARC growth.

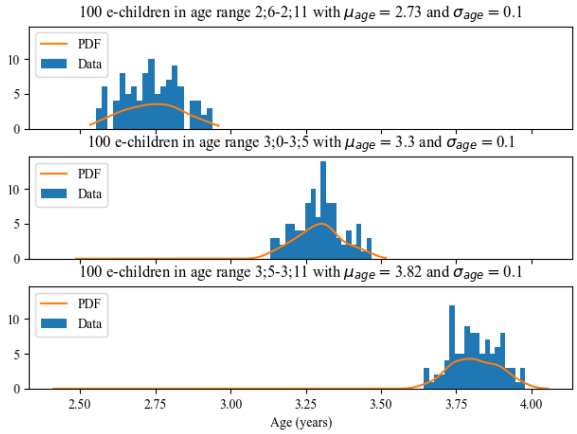
$$IARC_{\text{linear}}(U_i) = \begin{cases} 0 & U_i \leq -\frac{c}{m} \\ mU_i + c & -\frac{c}{m} \leq U_i \leq \frac{1-c}{m} \\ 1 & U_i \geq \frac{1-c}{m} \end{cases} \tag{3}$$

In addition to the $IARC_{linear}$ function, we also employ a logistic function implementation of IARC, $IARC_{logistic}$ as shown in Equation (4) bound by 0 and 1, with growth rate m and midpoint c. The logistic function exhibits an s-shaped (sigmoid) curve. For a sufficiently low m, the logistic function behaves almost linearly across the midpoint c and is asymptotic at the (0 and 1) endpoint values.

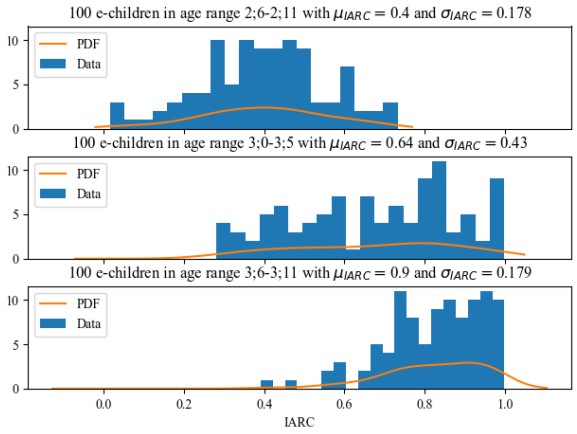
$$IARC_{logistic}(U_i) = \frac{1}{1 + e^{-m \times (U_i - c)}} \tag{4}$$

3.5 Simulation of a 100 e-children

Building on the research presented in Pearl and Sprouse (2021); Hart and Risley (1995, 2003), our estimate is that by age 5;0, a child from a professional-class background has been exposed to 10,054,267 utterances. To depict the variability among children noted in O&H's Experiment 2, we simulate 100 virtual children using a truncated Gaussian age distribution for each age category listed in Table 1, which specifies the minimum, maximum, and mean ages. Constructing a truncated Gaussian demands parameters such as range, mean (μ), and standard deviation (σ). O&H provide age ranges and mean ages (μ_{age}) in Table 1, but omit standard deviations for each group (σ_{age}). We approximate these standard deviations (σ_{age}) as 0.1 for all age groups based on available age ranges. According to Gleitman et al. (1984), imperatives make up about 16% of the language input a child receives until age 2. Earlier, Newport et al. (1977) estimated an 18% imperative usage beyond age 2. With $IARC = 0$, this reflects the expected level of



(a) Frequency histogram of ages.



(b) Frequency histogram of IARC values.

Figure 3: Performance of 100 e-children with a bin size of 20 and the Gaussian kernel estimation of the probability density function (*PDF*) across 3 age groups generated using a truncated Gaussian distribution emulating O&H.

	2;6-2;11	**3;0-3;5**	**3;6-3;11**
IARC Range	0-0.75	0.25-1	0.25-1
Tail end probability	(< 0.25) = 0.2	(> 0.75) = 0.4	(< 0.75) = 0.2
μ_{IARC}	0.4	0.64	0.9
σ_{IARC}	0.1785	0.43	0.179

Table 3: Table depicting the calculation of standard deviation of the Gaussian distribution of IARC parameter over age ranges.

noise (imperatives misunderstood as declaratives) encountered by the learner. Our simulations modulate the IARC parameter following Equations (3) or (4), assuming an imperative exposure rate of 16% up to age 2 (approximately 3,566,210 utterances) and 18% thereafter.

Similar to the age data, O&H do not provide the standard deviation of IARC (σ_{IARC}) for the 3 age groups. However, O&H do provide some additional distributional data which can be used to estimate the standard deviation. The O&H IARC data, presented in Table 2, has been recast as distributional metrics in Table 3. We infer the tail end probabilities of the IARC distribution from Table 2 in order to calculate the standard deviation (σ_{IARC}) of each age group. We observe that for ages 2;6-2;11, 20% of the children correctly interpret less than 2 out of 8 imperatives (IARC value less than 0.25), i.e., the probability that IARC is less than 0.25, $P(IARC) < 0.25$, is 0.2 (20% of the children). In addition, the mean IARC (μ_{IARC}) of this age range is 0.4 as reported in Section 2.1. With this tail end probability and the mean (μ_{IARC}), the standard deviation of the Gaussian distribution of the children's IARC value (σ_{IARC}) for the age range 2;6-2;11 was estimated to be 0.1785. The tail end probabilities for the other age groups were similarly inferred[7] and the standard deviations of all three age ranges are calculated and compiled in Table 3.

Using the parameters discussed above, a truncated Gaussian distribution in Scipy was used to generate an age distribution and an IARC distribution using two growth functions — $IARC_{linear}$ and $IARC_{logistic}$, of a 100 e-children across 3 age groups as depicted in Figures 3a and 3b respectively. After a 100 age and IARC values for each of the three age groups were generated, we sort the ages and the IARC values within each group. To approximate the longitudinal development of the IARC value for each e-child in the pool of a 100 e-children, we generate three $(IARC, age)$ pairs for each e-child, one from each age group, using the sorted IARC and age values. The first $(IARC, age)$ value in each of the three lists is used to generate e-child 1, the second three $(IARC, age)$ pairs are used to generate e-child 2, etc. Using these three $(IARC, age)$ pairs for each e-child, the optimal parameters for the two growth functions — $IARC_{linear}$ and $IARC_{logistic}$ — were calculated as outlined previously in this section. We then proceed to simulate the acquisition of the NS parameter for each of the resulting 100 e-children.

[7] For the middle age group, the right tail was used rather than the left.

Algorithm 2 Simulation of one e-child incorporating IARC.

G_{targ} is the target grammar $IARC$ is the probability of interpreting an imperative sentence correctly as an imperative
m and c are the optimal parameters for $IARC$ growth
$G_{targ} \leftarrow$ CoLAG English, i.e., 0001001100011
$IARC \leftarrow 0$
$num\_sentences \leftarrow$ 10,054,267, i.e., cumulative utterances by age 5;0
for *i in range(num_sentences)* **do**
  calculate IARC using m and c accordingly **if** $i < 3,566,210$, *i.e.,* $age < 2;0$ **then**
    $s \leftarrow$ sentence from the target language with 16 percent probability of being an imperative
  end
  else
    $s \leftarrow$ sentence from the target language with 18 percent probability of being an imperative
  end
  if s *is imperative* **then**
    with probability of, 1-IARC, interpret s as a declarative
  end
  Run SSVL on s
end

We conducted 2 experiments with a pool of 100 e-children employing the SSVL acquiring CoLAG English with 2 growth functions $IARC_{linear}$ and $IARC_{logistic}$. The e-children were generated according to the methodology described above. The simulations used an aggressive rate (R) of 2×10^{-4} and a conservative rate (r) of 5×10^{-6}. [8] The NS parameter weight / confidence values of these 100 e-children over time are plotted on the y-axis of the graphs presented in Figures 4a and 4b with the x-axis representing the number of cumulative utterances encountered. To show the variation of the e-children, all 100 are plotted with the fastest, the slowest, and the median e-child, in terms of convergence speed, demarcated.

4 Results

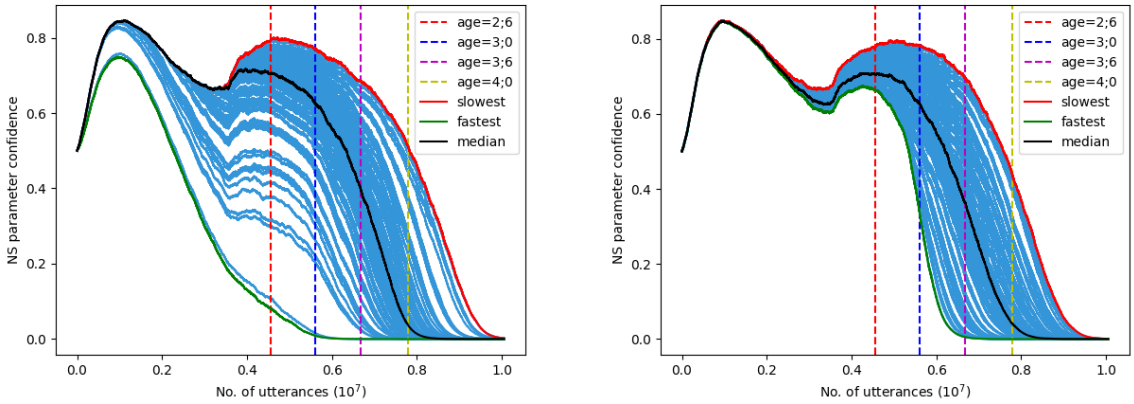
In the work of Orfitelli and Hyams (2012), a significant empirical discovery regarding developmental constraints is presented, specifically focusing on the differential interpretation of Null Subject (NS) sentences between adults and children. The NS stage arises from an intricate interplay of grammatical and performance elements. The objective of the study's simulations is to replicate this early-stage developmental constraint and the ensuing partial learning observed over time. The objective was to create electronic children, or e-children, whose linguistic development could accurately reflect the longitudinal findings of O&H. The simulations were conducted using a specifically adapted variational learning model (SSVL) incorporating superset and subset language frameworks, alongside two distinct models of IARC growth ($IARC_{linear}$ and $IARC_{logistic}$). The experiments with the SSVL model (illustrated in Figures 4a and 4b) reveal a particular behavior of the Null Subject parameter: it begins at an initial value of 0.5, then swiftly ascends to approximately 0.8, before subsequently declining, which mirrors the observed decrease in the employment of null subjects among English-speaking children. A minor resurgence occurs around age 2;0 due to a simulated increase in imperative sentence exposure experienced by an e-child, as elaborated in Section 3.5. This phenomenon is supported by findings from two distinct studies on imperatives directed at young children, one before the age of 2;0 and the other thereafter. Furthermore, we also performed simulations of the SSVL model in a noiseless setting within the CoLAG environment acquiring NS-English. Under noiseless conditions, SSVL demonstrates that the NS parameter promptly converges well before the e-children reach 2;0, the age traditionally associated with the onset of the NS stage.

5 Summary and Discussion

Orfitelli and Hyams (2012) observe that young English-speaking children often misinterpret (subjectless) imperative utterances as declaratives (e.g., *Play with blocks.*), which could potentially lead them to initially acquire an NS grammar. The present study computationally models the findings of Orfitelli and Hyams (2012). More generally, it establishes a framework for simulating a developmental *'performance parameter'* and its influence on acquisition. The performance parameter relevant to Orfitelli and Hyams (2012) and the computational work reported here we coin the *Illocution Ambiguity Resolution Coefficient* (IARC) — a measure of a child's ability to correctly disambiguate between imperative and declarative illocutionary force in utterances without a subject.

We computationally model the performance parameter IARC, based on empirical data from Orfitelli and Hyams (2012), and study its effect during acquisition of the NS syntactic parameter. Employing a modified version of the *Variational Learner* (VL, Yang 2002a, 2012), we simulate the change over time in the confidence value associated with

[8] The choice of learning rates was derived through trial and error.



(a) $IARC_{linear}$ growth function

(b) $IARC_{logistic}$ growth function

Figure 4: The SSVL employing the with learning rates $R = 2 \times 10^{-4}$, and $r = 5 \times 10^{-6}$ for 100 e-children. The fastest, the slowest and the median e-children, in terms of convergence speed, are highlighted.

the NS parameter in simulated 'e-children' acquiring an English-like language in an artificial language domain (Sakas and Fodor, 2011). The VL cannot reliably learn languages in superset/subset relationships Sakas et al. (2017), which is critical to modeling the acquisition of the NS parameter. To employ the VL paradigm in this context, we develop the *Superset/Subset Variational Learner* (SSVL) — a version of the VL that can effectively distinguish superset and subset grammars and successfully acquire them.

Simulating 100 SSVL e-children employing two growth functions of IARC, we observe that the IARC parameter's development over time affects each growth function in a similar fashion: Imperfect learning of the NS parameter early on, corrected later, converging on the obligatory-subject target grammar. Based on the psycholinguistic data presented in Orfitelli and Hyams (2012), one would expect to see an adjustment in the English-speaking child's grammar away from an NS grammar, as children grow to interpret subjectless imperative sentences correctly as imperatives (as modeled by the IARC parameter). The simulations conducted in this study reflect this trajectory of the NS parameter, which supports the conjecture presented in Orfitelli and Hyams (2012) — that the misinterpretation of subjectless imperatives is indeed a likely contributor to a child's Null Subject (NS) stage.

References

Robert C. Berwick. 1985. *The Acquisition of Syntactic Knowledge*. The MIT Press.

Lois Bloom. 1970. *Language Development: Form and Function in Emerging Grammars*. MIT Press.

Paul Bloom. 1990. Subjectless sentences in child language. *Linguistic Inquiry*, 21(4):491–504.

Robert R. Bush and Frederick Mosteller. 1955. *Stochastic Models for Learning*. Wiley, New York, NY.

Noam Chomsky. 1955. *The Logical Structure of Linguistic Theory*. Plenum Press and University of Chicago Press., New York, NY and Chicago, IL.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Noam Chomsky. 1981. *Lectures on government and binding*. Foris, Dordrecht, Netherlands.

Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York, NY.

Robin Clark. 1992. The selection of syntactic knowledge. *Language Acquisition*, 2(2):83–149.

Robin Clark and Ian Roberts. 1993. A computational model of language learnability and language change. *Linguistic Inquiry*, 24(2):299–345.

Stephen Crain and Janet Dean Fodor. 1993. The acquisition of structural restrictions on anaphora. In *Language and Cognition: A Developmental Perspective*, volume 16, pages 141–171. Ablex.

Stephen Crain and Cecile McKee. 1985. Acquisition of structural restrictions on anaphora. In *NELS*, volume 16, pages 94–110. University of Massachusetts, Graduate Linguistic Student Association.

Katherine Finn Davis, Kathy P. Parker, and Gary L. Montgomery. 2004. Sleep in infants and young children: Part one: normal sleep. *Journal of Pediatric Health Care*, 18(2):65–71.

Viviane Déprez and Amy Pierce. 1993. Negation and functional projections in early grammar. *Linguistic Inquiry*, 24(1):25–67.

Janet Dean Fodor. 1998. Unambiguous triggers. *Linguistic Inquiry*, 29(1):1–36. Type: Journal Article.

Janet Dean Fodor and William G. Sakas. 2005. The subset principle in syntax: costs of compliance. *Journal of Linguistics*, 41(3):513–569.

Lila R Gleitman, Elissa L Newport, and Henry Gleitman. 1984. The current status of the motherese hypothesis. *Journal of child language*, 11(1):43–79.

E Mark Gold. 1967. Language identification in the limit. *Information and Control*, 10(5):447–474.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Betty Hart and Todd R Risley. 2003. The early catastrophe: The 30 million word gap by age 3. *American educator*, 27(1):4–9.

Katherine Howitt, Soumik Dey, and William Gregory Sakas. 2021. Gradual syntactic triggering: The gradient parameter hypothesis. *Language Acquisition*, 28(1):65–96.

Eric H Lenneberg. 1967. The biological foundations of language. *Hospital Practice*, 2(12):59–67.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. Third Edition*. Lawrence Erlbaum Associates, Mahwah, NJ.

M. Rita Manzini and Kenneth Wexler. 1987. Parameters, binding theory, and learnability. *Linguistic Inquiry*, 18(3):413–444.

Elissa L. Newport, Henry Gleitman, and Lila R. Gleitman. 1977. Mother, I'd rather do it myself: Some effects and non effects of maternal speech style. In Catherine Elizabeth Snow and Charles Albert Ferguson, editors, *The acquisition of syntax*, page 109–149. Cambridge University Press.

Robyn Orfitelli and Nina Hyams. 2008. An experimental study of children's comprehension of null subjects: Implications for grammatical/performance accounts. In *Proceedings of the 32nd Annual BUCLD*, volume 2, pages 335–346.

Robyn Orfitelli and Nina Hyams. 2012. Children's grammar of null subjects: Evidence from comprehension. *Linguistic Inquiry*, 43(4):563–590.

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

Wilder Penfield and Lamar Roberts. 1959. *Speech and brain mechanisms*. Princeton University Press.

Luigi Rizzi. 2005a. Grammatically-based target-inconsistencies in child language. In *Proceedings of the Inaugural Conference of GALANA*. UCONN/MIT Working papers in Linguistics.

Luigi Rizzi. 2005b. On the grammatical basis of language development: A case study. In G. Cinque and R. Kayne, editors, *The Oxford handbook of comparative syntax*, pages 70–109. Oxford University Press.

William Gregory Sakas. 2003. A word-order database for testing computational models of language acquisition. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 415–422, Sapporo, Japan. Association for Computational Linguistics.

William Gregory Sakas and Janet Dean Fodor. 2011. Generating CoLAG languages using the "supergrammar.". Technical report.

William Gregory Sakas and Janet Dean Fodor. 2012. Disambiguating syntactic triggers. *Language Acquisition*, 19(2):83–143.

William Gregory Sakas, Charles Yang, and Robert Berwick. 2017. Parameter setting is feasible. *Linguistic Analysis*, 41:391–408.

Kenneth Jerold Straus. 2008. *Validations of a Probabilistic Model of Language Learning*. Ph.D. thesis, Department of Mathematics, Northeastern University.

Virginia Valian. 1990. Null subjects: A problem for parameter-setting models of language acquisition. *Cognition*, 35(2):105–122.

Virginia Valian. 1991. Syntactic subjects in the early speech of american and italian children. *Cognition*, 40(1):21–81.

Leslie Gabriel Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Qi Wang, Diane Lillo-Martin, Catherine T Best, and Andrea Levitt. 1992. Null subject versus null object: Some evidence from the acquisition of chinese and english. *Language acquisition*, 2(3):221–254.

Charles Yang. 2002a. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford, UK.

Charles Yang. 2002b. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford, UK.

Charles Yang. 2012. Computational models of syntactic acquisition. *WIREs Cognitive Science*, 3(2):205–213.

A Appendix

A.1 Convergence

Convergence is the learner's arrival at a final grammar hypothesis (G_{targ}). The final grammar hypothesis should license nearly all utterances of the target language and generate the same set of sentences. Under standard learnability assumptions, convergence is defined as arriving at a static grammar, i.e., one that will never change within a finite amount of time after entertaining a series of grammar hypotheses — Gold (1967), c.f., PAC-learning, Valiant (1984).

Integrating finiteness into a criterion of success is desirable in terms of formal learnability theory, and from an empirical standpoint — developmental psycholinguistic studies have established a period during which language learning occurs rapidly and apparently effortlessly. After this *critical period* (Penfield and Roberts, 1959; Lenneberg, 1967), the learner achieves a state of maturity with less plasticity in terms of language development (i.e., the learner converges on an adult grammar).

The implementation of this *finiteness criterion* varies between studies. For example, in Sakas et al. (2017) the criterion of successful convergence for the variational learner was a parametric weight threshold of 0.02 from the target parameter setting for each parameter, and in the case that the threshold was not met, the simulations were stopped after an e-child encountered 2 million utterances. Whereas, for the No-Defaults Learner in Howitt et al. (2021), simulations ended after an ad hoc number of sentences (500,000) were encountered by an e-child.

Pearl and Sprouse (2021, Appendix A, Table 9), estimate the number of sentences a real child hears between 2;4 and 5;0. They assume learning starts at 2;4 and calculated that from 28 months to 5 years a child from a professional family hears roughly 5,658,535 sentences. This calculation was based on Hart and Risley (1995, 2003), who provide data on how many sentences professional class parents speak to their children and Davis et al. (2004) who provide the average total daily sleep hours for children. In our case, however, we assume acquisition of the NS parameter starts at birth and estimate the number of sentences from birth to 5;0. We used Davis et al. (2004, Figure 1), which plots daytime and nighttime sleeping hours to plot total waking and total sleeping hours by age, see Figure 5.

Using the data presented in Figure 5, we estimate the number of sentences a child hears from birth to age 5;0. In order to develop the relevant calculations, we adopted three assumptions:

1. The number of waking hours of a child at age 1 month is almost the same as at birth.

2. The number of utterances per hour spoken by a parent to a child is uniform across all ages, i.e., 487 (Hart and Risley, 1995, 2003).[9]

3. The increase in waking hours across age intervals is linear.

When presenting our calculations, we employ the following notation. The *age period* (a_i) is the difference in years, between two points delineating a specified *age range* (i). The *daily waking hours* (h_i^1 to h_i^2) are the waking hours at the two endpoints of age range i. The *total waking hours* of a child in age range i is represented by H_i. *Total utterances* (u_i) is the total number of utterances heard by the child in age range i while the *cumulative utterances* (U_i) is the total number of utterances heard by a child from birth to the last date of age range i.

We now turn to how we calculate some of these variables. To calculate total utterances in age range i (u_i), and subsequently cumulative utterances by the end of age range i (U_i), we must first calculate the total waking hours at that age range (H_i). Figure 5 gives the number of waking hours at specific ages. Assuming the growth of waking hours between any two adjacent ages is linear (Assumption 3) — to calculate the total waking hours between two adjacent ages, we compute the area under the straight "line" of growth between the two age intervals and multiply the area by the number of days in a year (365), see Equation (5).

$$H_i = \frac{(h_i^1 + h_i^2)}{2} \times a_i \times 365 \tag{5}$$

The total utterances at age range i (u_i) is then derived, under Assumption (2) by Equation (6):

$$u_i = \lceil H_i \times 487 \rceil \tag{6}$$

Finally, we can then calculate the cumulative utterances at the end of age range i (U_i) using Equation (7):

$$U_i = U_i + U_{i-1} \tag{7}$$

[9] Pearl and Sprouse (2021) make a similar assumption.

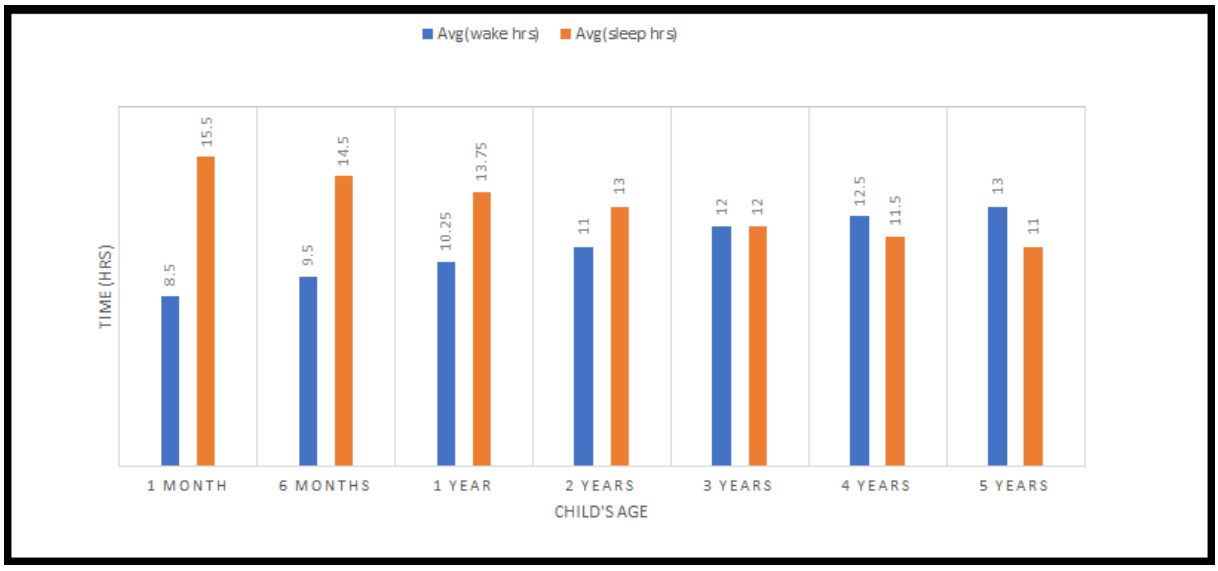


Figure 5: Average total daily sleep and waking hours for infants and young children. Data is taken from Davis et al. (2004).

age range (i)	**0;0 to 0;6**	**0;6 to 1;0**	**1;0 to 2;0**	**2;0 to 3;0**	**3;0 to 4;0**	**4;0 to 5;0**
age period (a_i)	0.5	0.5	1	1	1	1
daily waking hours (h_i^1 to h_i^2)	8.5-9.5	9.5-10.25	10.25-11	11-12	12-12.5	12.5-13
total waking hours (H_i)	1,642.5	1,802.19	3,878.13	4,197.5	4,471.25	4,653.75
total utterances (u_i)	799,898	877,665	1,888,647	2,044,183	2,177,499	2,266,376
cumulative utterances (U_i)	799,898	1,677,563	3,566,210	5,610,392	7,787,891	10,054,267

Table 4: Estimation of number of utterances encountered over different age ranges of child language acquisition.

The results of these calculations are presented in Table 4. Following Pearl and Sprouse (2021), we take the *stopping point* for our simulated e-children to be 5;0. The number of cumulative utterances at 5;0 per our calculations is 10,054,267.

We can also approximate the number of cumulative utterances heard by a child at any given age. For example, to calculate the utterances heard by a child at age 3.3 years, we first need to approximate the waking hours at that age. The difference between the number of waking hours between ages 3;0 (12 waking hours) and 4;0 (12.5 waking hours) is 0.5 hours. Since we assume linear growth, we can approximate the number of waking hours at age 3.3 years: $12.15 = 12 + (0.3 * 0.5)$. From Table 4, we know that the number of cumulative utterances at age 3 years is 5,610,392. The total utterances a child hears between 3 years and 3.3 years can be calculated according to Equations (5) and (6), as is illustrated in (8):

$$\begin{aligned} 1,322.2 &= \frac{12 + 12.15}{2} \times 0.3 \times 365 \\ 643,918 &= \lceil 1,322.2 \times 487 \rceil \end{aligned} \tag{8}$$

The cumulative utterances heard by age 3.3 years can then be calculated using Equation (7): $6,254,310 = 643,918 + 5,610,392$.

A.2 CoLAG domain details

Thirteen syntactic parameters were used to generate the languages and derivations in CoLAG (see Table 5). The target parameter values of CoLAG

Parameter List			
Parameter Name	Abbrev	Target Value = 0.0	Target Value =1.0
Subject Position	(SP)	Initial	Final
Headedness in IP	(HIP)	Initial	Final
Headedness in CP	(HCP)	Initial	Final
Optional Topic	(OpT)	Obligatory Topic	Optional Topic
Null Subject	(NS)	No Null Subject	Optional Null Subject
Null Topic	(NT)	No Null Topic	Optional Null Topic
Wh-Movement	(WhM)	Wh-Insitu	Obligatory Wh Movement
Preposition Stranding	(PI)	Obligatory Pied Piping	Optional Preposition Stranding
Topic Marking	(TM)	No Topic Marking	Obligatory Topic Marking
V to I Movemnt	(VtoI)	No VtoI Movement	Obligatory VtoI Movement
I to C Movement	(ItoC)	No ItoC Movement	Obligatory ItoC Movement
Affix Hopping	(AH)	No Affix Hopping	Affix Hopping
Question Inversion	(QInv)	No QInversion	Obligatory QInversion

Table 5: The 13 CoLAG parameters and their corresponding target values.

English are: 0001001100011 which corresponds from left to right, the values of the thirteen parameters in Table 5 from top to bottom. CoLAG English has word order patterns made up of the following lexical tokens: *S*, *01*, *02*, *03*, *P*, *Adv*, *Aux*, *Verb*, *not*, and *never*. These tokens correspond to subject, direct object, indirect object, object of a preposition, preposition, adverb, auxiliary, main verb, not and never respectively. CoLAG sentence patterns also have an overt (audible by e-children) illocutionary force feature: *Q*, *DEC* and *IMP* for questions, declaratives and imperatives respectively. An example English pattern in CoLAG is: *S Aux V O1 [DEC]* which might correspond to the natural language sentence: '*The little dragon is breaking the wall.*'.

CoLAG English has 360 distinct sentence patterns, 180 declaratives, 36 imperatives, and 144 questions. The Null Subject (NS) parameter is the parameter of interest here. If a CoLAG language is generated with NS=0 (e.g., CoLAG English), then every declarative and question has an overt subject. If NS=1, two versions of an utterance are generated, one with a subject and one without. The simulation studies detailed in this study present declaratives, questions, and imperatives to an e-child immersed in a CoLAG English-like language. Declaratives and questions are presented with overt subjects in CoLAG English. In CoLAG, imperative word orders universally do not have overt subjects.

A.3 Additional Algorithms

Algorithm 3 Variational Learner reward only.

for *each* w_i *in* W **do**
  set w_i to 0.5.
end
for *each input sentence s* **do**
  for *i in range(n)* **do**
    with probability w_i, parameter value $p_i^v \leftarrow 1$ with probability $1 - w_i$, parameter value $p_i^v \leftarrow 0$
  end
  G_{curr} = [p_1^v, ... ,p_n^v]
  if G_{curr} *can parse s* **then**
    for w_i *in* W **do**
      adjust w_i towards p_i^v using Equation (1) or (2);
    end
end

Algorithm 4 Simulating the TVJ experiment for a 100 e-children

IARC-list$_1$ ← sorted distribution of IARC for a 100 e-children of ages 2;6-2;11
IARC-list$_2$ ← sorted distribution of IARC for a 100 e-children of ages 3;0-3;5
IARC-list$_3$ ← sorted distribution of IARC for a 100 e-children of ages 3;6-3;11
age-list$_1$ ← sorted distribution of ages for a 100 e-children of ages 2;6-2;11
age-list$_2$ ← sorted distribution of ages for a 100 e-children of ages 3;0-3;5
age-list$_3$ ← sorted distribution of ages for a 100 e-children of ages 3;6-3;11
for *i in range (0 to 100)* **do**
  $IARC_1$ ← IARC-list$_1$[i]
  $IARC_2$ ← IARC-ist$_2$[i]
  $IARC_3$ ← IARC-list$_3$[i]
  age_1 ← age-list$_1$[i]
  age_2 ← age-list$_2$[i]
  age_3 ← age-list$_3$[i]
  Calculate optimal m and c using ($IARC_1$, age_1), ($IARC_2$, age_2), ($IARC_3$, age_3)
  Run Algorithm 2 with optimal m and c
end

A is for *a*-generics: Predicate Collectivity in Generic Constructions

Carlotta Marianna Cascino
Département d'Études Cognitives, École Normale Supérieure, Paris, France
Department of Psychology, Princeton University, Princeton, NJ, USA
carlotta.marianna.cascino@ens.psl.eu

Abstract

Generic statements like *A dog has four legs* are central to encode general knowledge. Yet their form–meaning mapping remains elusive. Some predicates sound natural with indefinite singulars (*a*-generics), while others require the definite article (*the*-generics) or the bare plural (bare-plural generics). For instance, why do we say *The computer revolutionized education* but not *A computer revolutionized education*? We propose a construction-based account explaining why not all generic statements are created equal. Prior accounts invoke semantic notions like kind-reference, stage-levelness, or accidental generalization, but offer no unified explanation. This paper introduces a new explanatory dimension: predicate collectivity level, i.e. whether the predicate applies to each member of a group or to the whole group as a unit (without necessarily applying to each of its members individually). Using two preregistered acceptability experiments we show that *a*-generics, unlike *the*-generics and bare-plural generics, are dispreferred with collective predicates. The findings offer a functionally motivated, empirically supported account of morphosyntactic variation in genericity, providing a new entry point for Construction Grammar.

1 Introduction

We interact meaningfully in the world on the basis of our knowledge of categories. A key reason humans (and other animals) categorize entities is to predict how to interact with new instances. Instances of a category tend to share properties with other members of the same category and not share properties with members of competing categories (Rosch & Mervis, 1975).

One way humans explicitly inform others about properties of categories is by using certain linguistic constructions, regularly referred to as generic statements. An aspect of generic statements that has garnered a great deal of attention is that people are willing to endorse generic statements even when a property only holds of a minority of instances. For instance, most people agree with the statement in (1) (Pelletier & Asher, 1997; Leslie et al., 2011), even though only adult female ducks lay eggs.

(1) Ducks lay eggs.

Much interest in generic statements concerns this fact, which distinguishes generics from universally quantified statements (*All duck lay eggs*). As in example (1), generic categories are often expressed using a bare plural form and much work on generics focuses on this type of generic (e.g., Carlson, 1977; Cohen & Erteschik-Shir, 1997, 2002; Kiss, 1998; Nyugen, 2020).

However, generic expressions in English can be expressed in alternative ways as well. In particular, generic meaning can be expressed with the indefinite singular article (*a*) as in (2), which we refer to here as *a*-generics. A third way of expressing generic meaning involves the definite article (*the*) with a singular noun as in (3), which we refer to here as *the*-generics.

(2) *A*-generic: A duck lays eggs.

(3) *The*-generic: The duck lays eggs.

Languages rarely offer speakers a choice between constructions without the choice being meaningful. The choice of one construction over another may signal a different interpretation, context, register, or dialect (Humboldt, 1999; Clark, 1987; Goldberg, 1995). And in fact, when distinct generic constructions have been considered, researchers have posited some functional distinction or other between them.

In a comparison between *a*-generics and bare plurals, Cohen (2001) argued that *a*-generics must express a rule or a regulation. Bare plurals, instead, may either express the same type of mean-

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 109–119

ing, or simply describe the way things happen to be. Others have likewise evoked the idea that *a*-generics convey law-like, nonaccidental generalizations (Greenberg, 2003), expressing necessary ("analytic") properties (Lawler, 1973; Burton-Roberts, 1977).

Furthermore, it has been suggested that bare plural generics but not *a*-generics are compatible with conjunctions of predicates that refer to equally good, but mutually incompatible characteristic properties, none of which are satisfied by the majority of the kind, as in example (4) (Nickel, 2008; Kirkpatrick, 2022):

(4a) Computers were invented in the 20th century and perfected in the 21st century.

(4b) #A computer was invented in the 20th century and perfected in the 21st century.

It is not the case, indeed, that most individual computers were both invented in the 20th century and perfected in the 21st century.

A-generics have been claimed to be further restricted by disallowing "stage-level" predicates, which take stages of individuals as arguments (Condoravdi, 1994), as in (5):

(5a) Penguins are endangered.

(5b) #A penguin is endangered.

Guerrini (2025) has recently argued that some restrictions on *a*-generics—like not allowing for accidental generalizations (see example 6c)—stem from the claim that the singular indefinite form cannot denote a "kind," because kinds are inherently plural entities.

A second research direction concerns the distinction between *a*-generics and *the*-generics. For instance, Platteau (1980:121-122), suggesting that the basic principles of definite and indefinite reference are also applicable to generic NPs, claimed that indefinite generics "refer to a random element of a certain species", such that "the selected sample has the same default properties as all the other members of the species". On the other hand—they claim—definite generics refer to one definite entity, which is "the abstract representative of the species".

Later work by Krifka (1987) and Krifka et al. (1995) distinguished the functions of indefinite generics ("I-generics") and definite generics ("D-generics") as follows.[1] Definite generics can involve "kind" predicates (6a), i.e. predicates whose subject is a kind; dynamic predicates (6b), i.e. non-stative predicates (see also Heyer, 1985); or accidental properties (6c) (see also Lawler, 1973; Burton-Roberts, 1977; Cohen, 2001; Greenberg, 2002, 2003). On the other hand, none of these types of predicates is possible with indefinite generics (7a-c):

(6a) The lion is extinct.

(6b) The rat reached Australia in 1770.

(6c) The madrigal is popular.

(7a) ?A lion is extinct.

(7b) ?A rat reached Australia in 1770.

(7c) ?A madrigal is popular.

Krifka further argues that only indefinite generics can be applied to "kinds that are not well-established," providing the contrast in (8) (see also Carlson, 1977):

(8a) ?The lion with three legs is ferocious.

(8b) A lion with three legs is ferocious.

As for the forms, according to Krifka (1987), singular definites, plural definites and taxonomic[2] generics belong to the class of "D-generics", while singular indefinites belong to the class of "I-generics". Bare plurals and bare singular generics, instead, have none of the restrictions just mentioned (see also Krifka, 2003), occurring in both classes, as shown in the following examples:

(10a) Lions are extinct.

(10b) Bronze was invented before 2000 B.C.

(11a) Rats reached Australia in 1770.

(11b) Rice was introduced in East Africa some centuries ago.

(12a) Madrigals are popular.

(12b) Music is popular.

(12a) Lions with three legs are ferocious.

(12b) Gold which is hammered flat is precious.

Overall, within this system, D-Genericity has been analyzed as "reference to kinds, which is NP-oriented", i.e. dependent on the type of noun phrase (13a); and I-genericity has been analyzed as "default quantification" which has scope over the

[1] As will become clearer later, Krifka's categories of "D-generics" and "I-generics" are not to be equated with specific grammatical forms such as the definite singular or the indefinite singular. Rather, definite generic NPs and indefinite generic NPs merely serve as their most typical realizations (Krifka, 1987: 4).

[2] Taxonomic generics have been claimed to have themselves different forms, and to refer to subspecies of a kind (Galmiche, 1985), as in the following examples:

(9a) *One lion*, namely the Asian lion, is nearly extinct.

(9b) *This lion* (the Asian lion) is nearly extinct.

(9c) *The rice they grow in East Africa* needs little water.

A detailed treatment of their possible forms lies beyond the scope of this paper.

VP as well (Krifka, 1987), occurring in "characterizing sentences" (13b), i.e. generalizations over groups of particular episodes of facts (Krifka et al., 1995). Krifka et al.'s proposal also allows for kind-referring NPs to occur in characterizing sentences (13c), recognizing potential overlaps in both form and meaning between I-generics and D-generics:

(13a) The potato was first cultivated in South America.

(13b) A potato contains vitamin C, amino acids, protein and thiamine.

(13c) The potato is highly digestible.

In our work, instead, we distinguish three types of generic constructions, which we refer to simply as *a*-generics, *the*-generics and bare-plural generics. This differs from Krifka's use of the labels "I-generics" and "D-generics", because we presume that the morphology provides an invitation to identify functional categories, and our goal is to determine what those categories are.

As for experimental work, Driemel et al. (2025) presented cross-linguistic evidence based on an acceptability judgements study testing singular definite, singular indefinite, bare plural, and definite plural generic forms. Their results show that bare plurals are preferred in English and German for kind- and characterizing-level readings, while definite plurals dominate in Romance and Greek. Although from their graph it is possible to note that definite singulars are preferred over indefinite singulars for kind reference, the authors do not explicitly mention it.

We are unaware of other prior experimental work testing distinctions among generic morphosyntactic forms, with the exception of Fuellenbach et al. (2019), who hypothesized that *a*-generics prefer normative or essential predicates ("principled": e.g. *A fep has red wings*) rather than incidental predicates ("statistical": e.g. *A fep throws glow sticks*). In a two-alternative forced choice task, child and adult participants were first exposed to an image of a target novel animal (e.g., a *kevta*) followed by a statement of one of four types:

Kevtas / A kevta / The kevta / This kevta wears scarves.

Participants were then asked: *Which one of these is also a kevta?* Only one of the images contained the same novel animal with the predicated property (e.g., a kevta wearing a scarf). Of interest was whether participants would interpret the statement as generic, in which case other instances of the same category should also share the same property (e.g., wear a scarf). *The*-generics were instead predicted to lead to lower generalisability with statistically connected property, but not necessarily to higher generalisability with principled properties[3]. They also predicted that bare plural subjects would support both principled and statistical properties equally well. Results showed that participants were more likely to treat the statement as generic when the predicate was normative or essential (e.g., "has red wings") than when the property was incidental "throws glow sticks") regardless of the morphosyntactic form of the statement.

2 Hypotheses

Much prior work on genericity has focused on the semantic compatibility between generics and certain types of predicates. Building on this literature, we hypothesize that a key factor influencing the acceptability of different generic constructions lies in whether the predicate is construed as applying to an individual or to a group. This distinction amounts to the well-known contrast between distributive and collective predicates. A *distributive* predicate applies individually to each member or subset of a group (or parts of an entity), while a *collective* predicate applies to a group or entity as a unit, without necessarily applying to each of its members individually (Link, 1983; Landman, 1989; Champollion, 2020).

For instance, the quantifiers *each* and *every* require a distributed interpretation, while *all* allows for a collective interpretation. That is, the statements in (14a) describe some very strong children, while (14b) allows an interpretation in which the children acted as a group to raise the turkey.

(14a) *Distributive*: Each / Every child lifted a 100 pound turkey.

(14b) *Collective*: All the children lifted a 100 pound turkey.

We hypothesized that the critical distinction between *a*-generics and *the*-generics is similar. Since the indefinite singular determiner, *a*, evokes a single indefinite individual, *a*-generics require predicates that can be construed as applying to (most)

[3]The author claimed that this pattern is similar to the one predicted for *a*-generics, but they expected *the*-generics to be rated lower overall, due to their overall more restricted use.

any[4] individual of the category. The meaning of *a*-generics, is motivated, on this perspective, by the fact that the predicate applies to any randomly selected individual of a category.

The definite singular determiner, *the*, on the other hand, generally combines with identifiable, specific nouns rather than any randomly selected member of a group. Therefore *the*-generic interpretations cannot be motivated in the same way as *a*-generics. Instead, we hypothesized that *the*-generics predicate a property of a clearly identifiable group or kind.

Construction	*a*-generics	*the*-generics
Morphosyntactic Form	[*A* N'] VP	[*The* N'] VP
Functional Constraints	VP predicate is construed to apply to a randomly selected instance of the category N'	VP predicate is construed to apply to the category as a collective or group

Table 1: Two hypothesized generic constructions in English

To understand the claims in Table 1, consider the pairs of sentences in (15) and (16). As confirmed by norming, described in the next section, (15) involves a collective predicate and (16) a distributive predicate. We predict that collective predicates will be rated more acceptable with *the*-generics (e.g., 15a) than with *a*-generics (15b), while distributive predicates will be rated more acceptable with *a*-generics (e.g., 16a) than with *the*-generics (16b):

Predicate construed to apply to a collective:

(15a) The bee pollinates crops across the globe.>

(15b) A bee pollinates crops across the globe.

Predicate construed to apply to (randomly chosen) individuals:

(16a) A bee dies after it stings. >

(16b) The bee dies after it stings.

In this paper we test in two preregistered studies[5] the prediction that the distinction between collective and distributive readings impacts English speakers' preference for *the*- vs. *a*-generics. The derivations of this prediction can be schematized as follows:

i) ***a*-generics favor predicates that apply to randomly chosen individual instances of a category**: *a*-generics will be judged less acceptable when combined with properties that apply to a category construed as a collective.

ii) ***the*-generics favor predicates that apply to a specific, clearly identifiable category**: *the*-generics will be judged more acceptable when combined with properties that apply to a category construed as a collective.

iii) **bare plurals display neither restriction.**

Our proposal draws inspiration from Platteau's (1980) distinction between reference to a random element of a certain species and reference to the abstract representative of the species. We also take up Krifka's (1987) call for a functional distinction between types of genericity, but reinterpret it within a constructional perspective, associating functional distinctions with morphosyntactic distinctions. Note also that predicate collectivity is orthogonal to kind-reference in the sense described by Krifka (1987): while kindhood concerns the referential status of the NP, collectivity captures how the VP applies across instances.

In sum, previous accounts have identified a wide range of semantic constraints on *a*-generics. Rather than replacing earlier insights, our findings identify a new explanatory dimension, i.e. predicate collectivity, proposing a novel empirically-grounded and morphosyntactically-oriented perspective on genericity.

3 Experiment 1

To test these predictions empirically, we conducted two experiments. Experiment 1 focused on how predicate collectivity influences the acceptability of *a*- vs. *the*-generics across a range of naturalistic sentences.

3.1 Methods

Participants. 79 native English speakers were recruited via Prolific (47F, 32M; M = 38.2 yrs) to provide acceptability ratings. As planned, participants who failed to accurately rate acceptable fillers higher than unacceptable fillers were excluded from analyses (mean rating unacceptable fillers $\geq$ mean rating acceptable fillers). This proved a stringent criterion and 23 participants were excluded, resulting in 56 participants included in the analysis. Because of the high number of exclusions, we also

[4] We include "most any" here because of the well-known fact that generic statements need not hold of every single instance of a group to be judged felicitous (e.g. *A duck lays eggs*).

[5] Link to the preregistrations: https://aspredicted.org/hs5y-b6p7.pdf

ran the analyses on all participants. Results on the predicted interaction did not change and can be found in Appendix A.

Materials. We constructed 12 English predicates (verb phrases) and count nouns in subject position (e.g., [computer] has transformed education) (see Appendix B for items). Twenty-four stimuli were then created by instantiating each noun phrase in two versions: with an *a*-generic (e.g., *A computer has transformed education*), and with a *the*-generic (e.g., *The computer has transformed education*). We additionally included 6 filler items. Each filler was a generic sentence with a bare plural subject; 3 fillers were intended to be fully acceptable, while three others deliberately contained grammatical errors. The latter served as attention checks and exclusion criterion. To quantify the degree of collectivity vs distributivity, we normed each of the 12 predicates combined with bare plural subjects. For this, a separate group of 22 native English speakers was recruited via Prolific and paid for their time to perform a forced-choice task asking whether each sentence was about individuals or groups. An example is provided in Figure 1:

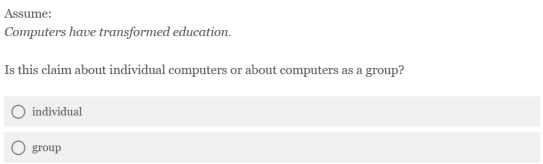


Figure 1: An example of the task in the norming experiment.

Participants who always responded with the same answer (n = 5) were excluded. The mean proportion of "group" responses from the remaining participants for each item was then used as predictor in the statistical analysis. This is a collectivity score, and ranged between 0 and 1. Predicates and their corresponding collectivity scores are reported in Appendix B.

Procedure. Each participant rated the acceptability of one version of each sentence (either an *a*-generic or *the*-generic), on a 7-point Likert scale, with generic type counterbalanced across participants. We further subdivided items so that each participant judged 6 target sentences: 3 *a*-generics and 3 *the*-generics, along with the 6 filler sentences. Items were presented in a randomized order for each participant. Instructions are provided in Figure 2.

An alternative design would have been to present participants with both versions of each sentence (with an *a*-generic and with a *the*-generic) and ask them to rank the two sentences of the pair for acceptability. A within-subjects setup of this kind typically affords greater statistical power by reducing variability across participants. However, exposure to one version would likely influence judgments of the other, turning the task into a relative rather than independent assessment. To avoid this, we adopted a design in which each participant was exposed to only one version of a given sentence, rating it independently of its alternative.

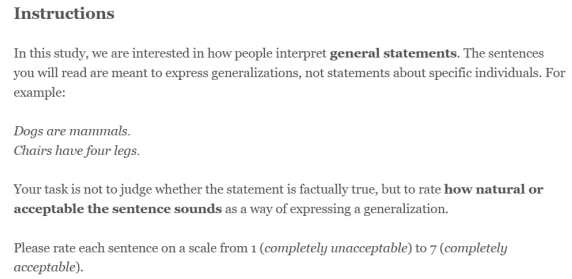


Figure 2: Instructions for Experiment 1.

3.2 Results

Results confirmed the hypotheses that *a*-generics and *the*-generics display different distributional patterns, and that *a*-generics were judged more acceptable when combined with a predicate that was more likely to be interpreted as applying to individuals (i.e., lower collectivity), rather than groups. This is shown in Figure 3.

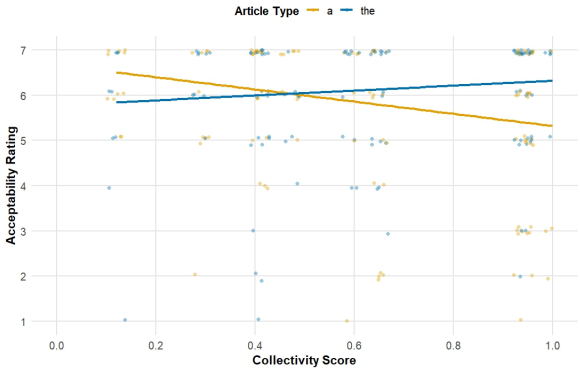


Figure 3: Acceptability ratings (1–7) as a function of collectivity scores (0–1) for *a*-generics and *the*-generics. Each point represents an individual response. Solid lines depict linear fits for each article type. Collectivity scores, normed separately, capture the degree to which a predicate was judged to refer to a collective vs. individual property. The figure illustrates that article acceptability interacts with collectivity, with *a*-generics associated with lower acceptability as collectivity increases.

This was confirmed with the planned cumulative link mixed-effects model (using the ordinal package in R [Christensen, 2019]. The output was raw acceptability ratings on the Likert scale. Article (*the* vs. *a*) and degree of collectivity were included as fixed interacting factors. Random intercepts and slopes for collectivity level were included by participant, while by-item random effects included intercepts and random slopes for article.

The model was fit to 342 observations (log-likelihood = –434.60; AIC = 901.21). The predicted interaction between article and predicate type was significant ($\beta = 2.45$, SE = 0.92, z = 2.66, p = .0077). Specifically, when the article was *a*, increasing collectivity scores significantly decreased acceptability ratings ($\beta = -1.68$, SE = 0.66, z = -2.56, p = .0105). A likelihood ratio test comparing the full model with article–collectivity interaction to a reduced additive model confirmed that including the interaction significantly improved model fit ($\chi^2(1) = 5.8367$, p = .0157), justifying its inclusion.

To examine this interaction more closely, we then fit separate models for *a*-generics and *the*-generics sentences[6]. For *a*-generics sentences, acceptability ratings decreased as collectivity increased, with a marginally significant negative effect of collectivity ($\beta = -1.81$, SE = 0.95, z = -1.9, p = .057). In contrast, for *the*-generics, collectivity showed a positive but non-significant effect on ratings ($\beta = 0.48$, SE = 0.79, z = 0.6, p = .546).

Taken together, these results indicate that the significant interaction observed in the full model is primarily driven by the sensitivity of *a*-generics to collectivity, whereas *the*-generics appear robust to this variation.

3.3 Discussion

Our results provide evidence that the morphosyntactic form of generic statements motivates their constraints. Specifically, as the predicate's collectivity score increased, *a*-generics significantly decreased in acceptability, as predicted by the constraint hypothesized in Table 1. As for *the*-generics, although the effect of collectivity did not reach statistical significance, the positive trend in the data warrants further investigation.

The contrast between collective and distributive predicates recalls Krifka et al.'s suggestion (1995; see also Krifka, 1987; Guerrini, 2025) that *a*-generics do not allow subjects that refer to kinds. The current proposal goes beyond this observation in several ways. First, we demonstrate that *a*-generics disprefer not only kind-level predicates (as suggested by Krifka et al., 1995), but collective predicates more broadly. This includes cases that do not involve reference to kinds *per se*. For instance, *a*-generics are rated significantly less acceptable with predicates such as *pollinates crops across the globe* or *hunts in packs* (e.g., *A bee pollinates crops across the globe*, *A wolf hunts in packs*), both of which attribute properties to the collective behaviour of a cateogry. Secondly, we show that not only do *a*-generics disprefer predicates that apply to groups, we positively characterize the type of interpretation *a*-generics prefer: *a*-generics prefer predicates that apply to a randomly selected instance of a category. Furthermore, we motivate why the *a*-generic construction patterns the way it does: the conventional referential profile associated with indefinite NPs in English helps explain its functional constraints in generic interpretation. As a result, we predict that languages with analogous morphosyntactic distinctions (e.g., indefinite singular vs. definite singular forms) will exhibit similar distributional tendencies, and that reversed patterns would be typologically rare or marked.

4 Experiment 2

While Experiment 1 confirmed our core prediction, it left open the behavior of bare-plural generics. Experiment 2 introduces this additional form to evaluate whether it patterns more like *the*-generics or *a*-generics in its sensitivity to predicate collectivity.

4.1 Methods

Participants. We recruited 116 native English speakers via Prolific (68F, 45M, 2NB; M=37yrs). As planned, participants whose mean rating of three ungrammatical catches was equal to or higher than the mean rating of grammatical bare plurals were excluded from analyses (n = 25, excluded). Reported analyses were therefore run on 91 participants.

Materials. The same *a*-generic and *the*-generic sentences used in Experiment 1 were included, now along with a bare plural generic as well (e.g., *Computers have transformed education*). Since bare-plural generics were a new condition, we re-

[6] In doing so, we had to drop the random slope for article by item due to convergence issues.

duced the 6 filler sentences in Experiment 1 to 3 ungrammatical catch trials (all in bare plural form). As in Experiment 1, these 3 sentences served as attention checks and exclusion criterion. The stimuli can be found in Appendix C.

Procedure. The procedure for this experiment was the same as in Experiment 1 (participants were asked to rate each sentence's acceptability on a 7-point Likert scale). Each participant saw one version of each sentence (either with *a*-generic, *the*-generic or bare plural generic), with article type counterbalanced across participants. We further subdivided items, so that each participant judged 6 target sentences—2 *the*-generics, 2 *a*-generics and 2 bare-plural generics, from one of six lists, assigned randomly, along with the 3 catch trials. Items were presented in a randomized order for each participant.

4.2 Results

Results again confirmed that the perceived naturalness of generic noun phrases is modulated by the collective properties of the predicate. As shown in Figure 4, *a*-generics received the highest acceptability ratings when combined with predicates that were less collective, while receiving the lowest acceptability with predicates that were more collective. Bare-plural generics and *the*-generics trended toward greater acceptability with more collective predicates, though this effect did not reach significance in isolation.

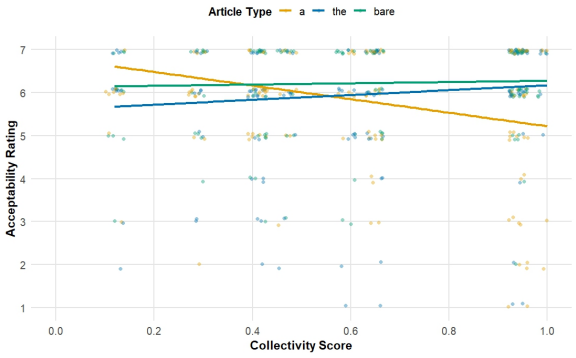


Figure 4: Acceptability ratings (1–7) as a function of collectivity scores (0–1) for *a*-generics, *the*-generics and bare generics. Each point represents an individual response. Solid lines depict linear fits for each article type. Collectivity scores, normed separately, capture the degree to which a predicate was judged to refer to a collective vs. individual property. The figure illustrates that article acceptability interacts with collectivity, with *a*-generics associated with lower acceptability as collectivity increases.

This was confirmed with the planned cumulative link mixed-effects model (fitted using the ordinal package in R; Christensen, 2019). The output was raw acceptability ratings on the Likert scale. Article (definite singular vs. indefinite singular vs. bare plural) and degree of collectivity were included as fixed interacting factors. Random intercepts and slopes for collectivity were included by participant, while by-item random effects included intercepts and random slopes for article. The model was fit to 546 observations (log-likelihood = –682.52; AIC = 1407.04). The predicted interaction between article type and predicate collectivity was statistically significant. While *a*-generics showed a negative effect of collectivity on acceptability ($\beta = -2.03$, SE = 0.9482, z = -2.137, p = .033), both bare and *the*-generics showed significantly more positive slopes compared to *a*-generics ($\beta = +3.02$, SE = 1.09, z = 2.76, p = .006, and $\beta = +3.06$, SE = 1.13, z = 2.71, p = .007, respectively), reversing the trend. Separate models for each article type[7] replicated what we saw in Experiment 1: *a*-generics showed a moderately significant decrease in acceptability as collectivity increased ($\beta = -1.62$, SE = 0.93, z = -1.735, p = .083). *The*-generics and bare plurals numerically trended in the opposite direction, but there was no significant effect of collectivity on acceptability for either *the*-generics ($\beta = +1.27$, SE = 0.80, z = 1.58, p = .114) or bare-generics ($\beta = +0.41$, SE = 0.69, z = 0.60, p = .551).

4.3 Discussion

This second experiment builds on Experiment 1 by introducing bare-plural generics, thereby allowing us to assess how they pattern with respect to predicate collectivity level. The results replicate the core finding from Experiment 1: *a*-generics decrease in acceptability as predicate collectivity increases, aligning with previous proposals that they are anchored in random instance interpretation. On the other hand, *the*-generics exhibit the opposite trend, albeit non-significantly when tested in isolation. Bare-plural generics show a positive trend similar to *the*-generics, suggesting they may prefer collective predicates more than *a*-generics. Crucially, although the positive effects of collectivity on *the*- and bare-plural generics did not reach significance in isolation, the interaction structure

[7] In fitting separate models for each article type, we had to drop the random slope for article by item and the random slope for collectivity level by participant, due to convergence issues.

of the model shows that their behavior is reliably distinct from that of *a*-generics.

This supports the broader claim that the morphosyntactic form of generic constructions modulates how it interacts with the properties of the predicate. This distinction helps motivate why languages differentiate morphosyntactic strategies for expressing genericity: each form carries its own functional constraints—whether tight or loose.

5 Conclusion

The findings presented in this paper provide the first experimental evidence that genericity is sensitive to constructional variation. By systematical comparisons, we demonstrate that different morphosyntactic forms are not interchangeable. Instead, each encodes distinct functional constraints motivated by more typical uses of the same forms. In particular, unlike *the*-generics and bare-plural generics, *a*-generics tend to combine with VPs predicating a property of any individual member of a category, but not of a collectivity. These findings support the Construction Grammar insight that it is *constructions* (i.e. form-meaning pairings), and not merely lexical items or semantic content, that shape interpretive possibilities.

Other factors may plausibly influence the choice between *a*-generics, *the*-generics and bare-plural generics. For instance, noun type (e.g. mass/count), register, or information structure might modulate acceptability judgments. Future research could explore how these factors interact with predicate collectivity level, possibly by extending the stimuli dataset. Future work may also investigate how morphosyntactic distinctions correlate with collectivity in other languages, and whether such patterns can be captured or induced in Large Language Models.

While preliminary, these findings lay the foundation for a broader empirical research agenda focused on genericity as a construction-sensitive phenomenon.

6 References

Bäck, A. (1996). Review of *Generische Kennzeichnungen*, by Gerhard Heyer (1987). *Noûs*, *30*(2), 276–281.

Bencini, G. M. L., & Goldberg, A. E. (2000). The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, *43*(4), 640–651.

Borik, O., & Espinal, M. T. (2015). Reference to kinds and to other generic expressions in Spanish: Definiteness and number. *The Linguistic Review*, *32*, 167–225.

Burton-Roberts, N. (1977). Generic sentences and analyticity. *Studies in Language*, *1*, 155–196.

Carlson, G. N. (1977). *Reference to kinds in English* (Doctoral dissertation). University of Massachusetts, Amherst.

Carlson, G. N. (1989). On the semantic composition of English generic sentences. In G. Chierchia, B. H. Partee, & R. Turner (Eds.), *Properties, Types and Meaning* (pp. 167–192). Dordrecht: Springer.

Carlson, G. N. (1995). Truth conditions of generic sentences: Two contrasting views. In G. N. Carlson & F. J. Pelletier (Eds.), *The Generic Book* (pp. 224–237). Chicago: University of Chicago Press.

Carlson, G. N. (2011). Genericity. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook* (Vol. II, pp. 1153–1185). Berlin: de Gruyter.

Champollion, L. (2020). Distributivity, collectivity and cumulativity. In L. Matthewson, C. Meier, H. Rullmann, & T. E. Zimmermann (Eds.), *Wiley's Companion to Semantics*.

Chierchia, G. (1995). Individual level predicates as inherent generics. In G. N. Carlson & F. J. Pelletier (Eds.), *The Generic Book* (pp. 176–224). Chicago: University of Chicago Press.

Chierchia, G. (1998). Reference to kinds across languages. *Natural Language Semantics*, *6*, 339–405.

Chierchia, G. (2022). 'People are fed up; don't mess with them.' Non-quantificational arguments and polarity reversals. *Journal of Semantics*, *39*, 475–521.

Clark, H. H. (1987). *Arenas of language use*. Chicago: University of Chicago Press.

Cohen, A., & Erteschik-Shir, N. (2002). Topic, focus, and the interpretation of bare plurals.

Natural Language Semantics, *10*(2), 125–165.

Cohen, A. (2001). On the generic use of indefinite singulars. *Journal of Semantics*, *18*(3), 183–209.

Condoravdi, C. (1994). *Descriptions in contexts* (Doctoral dissertation). Yale University.

Christensen, R. H. B. (2019). *ordinal — Regression Models for Ordinal Data* (R package version 2019.12-10). Retrieved from https://CRAN.R-project.org/package=ordinal

Dahl, O. (1985). *Tense and aspect systems*. Oxford: Blackwell.

Dayal, V. (2013). On the existential force of bare plurals across languages. In I. Caponigro & C. Cecchetto (Eds.), *From Grammar to Meaning* (pp. 49–80). Cambridge: Cambridge University Press.

Diesing, M. (1992). *Indefinites*. Cambridge, MA: MIT Press.

Dotlačil, J. (2010). *Anaphora and distributivity: A study of same, different, reciprocals and others* (Doctoral dissertation). Utrecht University.

Driemel, I., Hein, J., Carioti, D., Wünsch, J., Tsakali, V., Alexiadou, A., Sauerland, U., & Guasti, M. T. (2025). An experimental study on kind and generic readings across languages: Bare plural vs. definite plural. *Proceedings of the Amsterdam Colloquium*, 360–366.

Farkas, D. F., & de Swart, H. (2005). The generic article. In *Proceedings of the ESSLLI Workshop on Cross-linguistic Semantics*.

Fuellenbach, K. (2020). *A generic subject: The interplay of morphosyntax and the human conceptual system* (Doctoral dissertation). University of Oxford.

Fuellenbach, K., Gelman, S. A., Husband, E. M., & Cuneo, N. (2019). Generalising properties based on the morphosyntax of the subject: Generic and non-generic interpretations. *AMLaP 2019*, Moscow, Russia.

Galmiche, M. (1985). *Phrases, syntagmes et articles génériques*. Langage 79.

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

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

Greenberg, Y. (2002). Two types of quantificational modalized genericity, and the interpretation of bare plural and indefinite singular NPs. In *Proceedings of SALT 12* (pp. 104–123).

Greenberg, Y. (2003). *Manifestations of genericity*. New York: Routledge.

Guerrini, J. (2024). Revisiting kind predication. [Manuscript under revision].

Guerrini, J. (2025). English bare plurals and distributivity. In N. Webster et al. (Eds.), *Proceedings of the 41st West Coast Conference on Formal Linguistics* (pp. 179–183).

Heyer, G. (1985). Generic descriptions, default reasoning and typicality. *Theoretical Linguistics*.

Heyer, G. (1987). *Generische Kennzeichnungen: Zur Logik und Ontologie Generischer Bedeutung*. München: Philosophia Verlag.

Humboldt, W. von. (1999). *On language: On the diversity of human language construction and its influence on the mental development of the human species* (P. Heath, Trans.; M. Losonsky, Ed.). Cambridge: Cambridge University Press. (Original work published 1836)

Kirkpatrick, J. (2022). Generic conjunctivitis. *Linguistics and Philosophy*, *45*.

Kiss, K. E. (1998). On generic and existential bare plurals and the classification of predicates. In S. Rothstein (Ed.), *Events and Grammar* (pp. 145–162). Dordrecht: Kluwer.

Kratzer, A. (1995). Stage-level and individual-level predicates. In G. N. Carlson & F. J. Pelletier (Eds.), *The Generic Book* (pp. 125–176). Chicago: University of Chicago Press.

Krifka, M. (1987). An outline of genericity. *SNS-Bericht 87–23*, University of Tübingen.

Krifka, M. (1989). Nominal reference, temporal constitution and quantification in event semantics. In R. Bartsch, J. van Benthem, & P. van Emde Boas (Eds.), *Semantics and Contextual*

Expressions (pp. 75–115). Dordrecht: Foris.

Krifka, M., Pelletier, F. J., Carlson, G. N., Meulen, A., Link, G., & Chierchia, G. (1995). Genericity: An introduction. In G. N. Carlson & F. J. Pelletier (Eds.), *The Generic Book* (pp. 1–124). Chicago: University of Chicago Press.

Krifka, M. (2003). Bare NPs: Kind-referring, indefinites, both, or neither? In *Semantics and linguistic theory XIII*, 180-203.

Landman, F. (1989). Groups, I. *Linguistics and Philosophy*, *12*(5), 559–605.

Langacker, R. W. (1997). Generics and habituals. *Current Issues in Linguistic Theory*, *143*, 191–222.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Lawler, J. (1973). Studies in English generics. *University of Michigan Papers in Linguistics*.

Leslie, S. J., Khemlani, S., & Glucksberg, S. (2011). Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, *65*(1), 15–31.

Leslie, S. J. (2015). 'Hillary Clinton is the only man in the Obama Administration': Dual character concepts, generics, and gender. *Analytic Philosophy*, *56*(2), 111–141.

Leslie, S. J., & Lerner, A. (2022). Generic generalizations. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition). Retrieved from https://plato.stanford.edu/archives/fall2022/entries/generics/

Nickel, B. (2008). Generics and the ways of normality. *Linguistics and Philosophy*, 31, 629–648.

Nunberg, G., & Pan, C. (1975). Inferring quantification in generic sentences. In *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society (CLS 11)* (pp. 412–422).

Nyugen, A. (2020). The radical account of bare plural generics. *Philosophical Studies*, *177*, 1303–1331.

Ojeda, A. (1991). Definite descriptions and definite generics. *Linguistics and Philosophy*, *14*, 367–397.

Pelletier, F. J., & Asher, N. (1997). Generics and defaults. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of Logic and Language* (pp. 1125–1179). Cambridge, MA: MIT Press.

Platteau, F. (1980). Definite and indefinite generics. In J. van der Auwera (Ed.), *The Semantics of Determiners* (pp. 112–123). London: Croom Helm.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Scha, R. (1981). Distributive, collective and cumulative quantification. In J. Groenendijk, T. Janssen, & M. Stokhof (Eds.), *Formal Methods in the Study of Language*. Amsterdam: Mathematical Centre Tracts.

Wilkinson, K. (1995). The common noun kind. In G. N. Carlson & F. J. Pelletier (Eds.), *The Generic Book* (pp. 383–397). Chicago: University of Chicago Press.

Appendices

A Analysis without exclusions (Exp. 1)

For this analysis, we fitted the same model as the one with exclusions, with the only exception that we had to drop the random intercept for item due to convergence issues. The model was fit to 486 observations (log-likelihood = –633.68; AIC = 1297.36). The predicted interaction between article and predicate type was significant ($\beta = 2.04$, SE = 0.84, z = 2.42, p = .0153).

We then fitted separate models for *a*-generic and *the*-generic sentences. Due to convergence issues, we simplified the random-effects structure by dropping the random slope for article by item. *A*-generics sentences showed negative but non-significant effects of collectivity on acceptability ratings ($\beta = -0.75$, SE = 0.81, z = -0.92, p = .358). In contrast, for *the*-generics, collectivity showed a positive but non-significant effect on ratings ($\beta = 0.76$, SE = 0.65, z = 1.16, p = .243).

B Sentence Stimuli for Experiment 1

Item	TARGET SENTENCES	Collectivity Score
1	The wolf hunts in packs. A wolf hunts in packs.	0.94
2	The wolf sharpens its teeth on bones. A wolf sharpens its teeth on bones.	0.41
3	The bee pollinates crops across the globe. A bee pollinates crops across the globe.	0.65
4	The bee dies after it stings. A bee dies after it stings.	0.12
5	The airplane revolutionized global travel. An airplane revolutionized global travel.	1.00
6	The airplane lowers its gear before landing. An airplane lowers its gear before landing.	0.59
7	The computer has transformed education. A computer has transformed education.	0.94
8	The computer boots up in seconds. A computer boots up in seconds.	0.29
9	The elephant is the largest land animal. An elephant is the largest land animal.	0.65
10	The elephant flaps its ears to cool down. An elephant flaps its ears to cool down.	0.41
11	The microwave modernized home cooking. A microwave modernized home cooking.	0.94
12	The microwave heats food in minutes. A microwave heats food in minutes.	0.47
FILLER SENTENCES		
	Cats purr them when they are content.	
	Birds build nests in spring.	
	Students often study late before exams.	
	Phones distract to people during meetings.	
	Doctors help to patients managing pain.	
	Plants grow faster in sunlight.	

C Sentence Stimuli for Experiment 2

Item	TARGET SENTENCES	Collectivity Score
1	The wolf hunts in packs. A wolf hunts in packs. Wolves hunt in packs.	0.94
2	The wolf sharpens its teeth on bones. A wolf sharpens its teeth on bones. Wolves sharpen their teeth on bones.	0.41
3	The bee pollinates crops across the globe. A bee pollinates crops across the globe. Bees pollinate crops across the globe.	0.65
4	The bee dies after it stings. A bee dies after it stings. Bees die after they sting.	0.12
5	The airplane revolutionized global travel. An airplane revolutionized global travel. Airplanes revolutionized global travel.	1.00
6	The airplane lowers its gear before landing. An airplane lowers its gear before landing. Airplanes lower their gear before landing.	0.59
7	The computer has transformed education. A computer has transformed education. Computers have transformed education.	0.94
8	The computer boots up in seconds. A computer boots up in seconds. Computers boot up in seconds.	0.29
9	The elephant is the largest land animal. An elephant is the largest land animal. Elephants are the largest land animals.	0.65
10	The elephant flaps its ears to cool down. An elephant flaps its ears to cool down. Elephants flap their ears to cool down.	0.41
11	The microwave modernized home cooking. A microwave modernized home cooking. Microwaves modernized home cooking.	0.94
12	The microwave heats food in minutes. A microwave heats food in minutes. Microwaves heat food in minutes.	0.47
FILLER SENTENCES		
	Cats drink quickly milk.	
	Students study often late before exams.	
	Doctors help to patients managing pain.	

Rethinking Linguistic Structures as Dynamic Tensegrities

Remi van Trijp
Sony Computer Science Laboratories, Paris Lab
6 Rue Amyot
75005 Paris (France)
remi.vantrijp@sony.com

Abstract

Constructional approaches to language have evolved from rigid tree-based representations to framing constructions as flexible, multidimensional pairings of form and function. However, it remains unclear how to formalize this conceptual shift in ways that are both computationally scalable and scientifically insightful. This paper proposes *dynamic tensegrity* – a term derived from "tensile integrity" – as a novel architecture metaphor for modelling linguistic form. It argues that linguistic structure emerges from dynamically evolving networks of constraint-based tensions rather than fixed hierarchies. The paper explores the theoretical consequences of this view, supplemented with a proof-of-concept implementation in Fluid Construction Grammar, demonstrating how a tensegrity-inspired approach can support robustness and adaptivity in language processing.

1 Introduction

Since its conception in the 1980s, Construction Grammar has evolved from a bold challenger of the field's core-periphery distinction into one of the most widely adopted frameworks in contemporary linguistics. The **constructional idea** – that all linguistic knowledge can be described as pairings of form and function, called *constructions* – collapsed the traditional boundaries between rules and exceptions, and between lexicon and grammar (Fillmore, 1988, 1989; Fillmore et al., 1988).

As the field shifted from more traditional, static descriptions towards dynamic usage-based approaches (Bybee and Thompson, 2000; Langacker, 2000; Goldberg, 2006), the initial conception of constructions as constituent trees was replaced by a view of them as multidimensional structures (Fried and Östman, 2004; van Trijp, 2016; Goldberg, 2019). However, it has proven difficult to formalize this conceptual shift in a computationally scalable and scientifically interpretable way.

Most current analyses in construction grammar still rely on tree-like or slot-filler architectures inherited from earlier paradigms. While these models are useful for static descriptions of linguistic structure, they are not adapted for handling the fluidity and adaptivity required for usage-based models. In response, several researchers have begun to investigate the relevance of Large Language Models (LLMs; Goldberg, 2024; Piantadosi, 2024); but although LLMs undeniably offer new possibilities, their lack of explicit structures makes them difficult to interpret – particularly for formulating scientific generalizations about how constructions contribute to meaning-making in situated interactions.

This paper aims to complement this modeling landscape by offering an explicit, constraint-based account of linguistic structure that unifies fluidity and robustness, while remaining fully interpretable for the human researcher. More specifically, we propose *dynamic tensegrity* – a structural principle used in architecture, robotics, and some biological models to explain how systems maintain stability through distributed tension and compression – as a novel metaphor for describing linguistic form.

Rather than representing linguistic structure as a rigid hierarchy, we model it as an evolving network of interdependent constraints held in a dynamic equilibrium. In this view, constructions combine to build structures that self-stabilize through ongoing resolution of interdependent morphosyntactic, semantic and pragmatic constraints, much like tensegrity structures distribute mechanical forces across the system to preserve structural balance.

We explore the theoretical consequences of this reframing and present a proof-of-concept implementation in Fluid Construction Grammar (FCG Steels, 2004, 2011; van Trijp et al., 2022; Beuls and Van Eecke, 2026), an open-source computational platform explicitly designed for developing adaptive yet robust models of language processing.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 120–129
September 24, 2025,

2 The Case for Constructional Integrity

All complex systems – whether biological, architectural or computational – require components that can sustain their *integrity* under dynamic conditions. In linguistics, however, integrity has often been misinterpreted as rigidity: as something that must remain fixed or untouched. In reality, integrity is what enables a system to maintain structural coherence while remaining flexible enough to function under pressure. In this section, we argue that this systems-level insight requires us to rethink the nature of grammatical structure. We propose **constructional integrity** as a foundational principle that explains how language can be both stable and adaptive – capable of preserving meaningful structure while dynamically responding to the demands of communication.

2.1 Integrity Misunderstood

During the heydays of transformational syntax, Noam Chomsky (1970) famously introduced the "lexicalist hypothesis", proposing a strict separation between word formation (morphology) and sentence formation (syntax). According to this view, the morphological component produces lexical items as ready-made parts, which are then arranged into syntactic configurations by phrase structure rules and transformations. By enriching the lexicon and minimizing the burden on syntax, Chomsky aimed to advance his broader goal of developing a theory of Universal Grammar.

The Lexicalist Hypothesis influenced the field far beyond transformational syntax, and led to the formulation of the **Lexical Integrity Principle** (Wasow, 1977; Lapointe, 1980), which Haspelmath and Sims (2010, p. 203) define as follows:

> "Rules of syntax can refer/apply to entire words or the properties of entire words, but not to the internal parts of their words or their properties."

The Lexical Integrity Principle is committed to rigidity: words are treated as atomic units that can be rearranged but not internally modified. At first glance, this seems plausible. Words do exhibit a cohesiveness that larger structures seem to lack. Take the sentences in (1), from Goldberg (2006, p. 21), which preserve the same underlying argument structure – an agent (*Nina*) transferring a patient (*a dozen flowers*) to a recipient (*her mother*) – despite differences in surface order.

(1) a. Nina sent her mother a dozen flowers.
b. A dozen flowers, Nina sent her mother!

In lexicalist approaches, these argument structure relations are already determined in the meaning of the verb. Surface alternations are then explained through transformations of a shared deep structure (e.g. Haegeman, 1994), or through lexical rules that modify the verb's syntactic behaviour (Briscoe and Copestake, 1999).

Constructional approaches, however, take a different view. According to Goldberg (2006), the argument structure relations in both sentences are contributed not by the verb alone, but by the Ditransitive construction – a more abstract argument structure construction that expresses Caused-Transfer semantics. Word order differences are attributed to the interaction of this construction with others – such as the topicalization construction – to satisfy discourse-pragmatic needs. In this view, meaning and structure are not projected from verb-centered templates, but emerge from the dynamic composition of constructions in context.

More importantly, the constructional view does not treat the Ditransitive construction as a rigid, phrase-structural template. Instead, it assumes a high degree of structural flexibility: the construction can be used in various configurations, such as topicalization or clefting, while preserving its core semantic function of indicating who does what to whom.

This kind of flexibility requires a form of integrity that we observe in living systems as well: the ability to maintain functional coherence while adapting to functional pressures. A clear illustration is the biological cell. As Huang et al. (2006, p. 290) note, "death of both cells and whole organisms is characterized by a rapid increase in rigidity (*rigor mortis*), with a complete loss of the flexibility that dominates the living state. Thus, this unification of *robustness* with *flexibility*, both in terms of cell structure and behavior, is a hallmark of complex living systems."

Crucially, while the rigid principle of Lexical Integrity has already been shown to be empirically inadequate (Bruening, 2018), we will argue that this kind of dynamic integrity is not a property of the lexicon alone, but of *all* constructions more broadly – including those that handle argument structure, information packaging, and discourse-level coordination.

2.2 Constructional Integrity in Action

Let's illustrate Constructional Integrity through some examples, starting with (2), which shows ellipsis in the coordination between *hand-drawn* and *computer-drawn*:

(2) Do you prefer *hand- or computer-drawn* animation?

This example violates the Lexical Integrity Principle, which prohibits syntactic operations from tampering with the internal structure of words. Here, the coordination construction must elide the second component of *hand-drawn*.

From the perspective of Constructional Integrity, however, such cases are not exceptions anymore. Constructions are mappings between form and function, where the form itself plays a *diagnostic* role: it enables language users to detect which constructional knowledge to activate (Croft, 2001). Morphological constructions are typically recognizable by their specific arrangement of phonemes, while syntactic constructions – such as the English passive construction – are typically recognizable by surface patterns (e.g. auxiliary-*have* + *ed*-participle).

In most contexts, removing *drawn* from *hand-drawn* would indeed be detrimental: it would render the construction unrecognizable, disrupting its communicative function. In the coordinated phrase in (2), however, functional integrity is preserved. The structural "load" is shared with *computer-drawn*, which enables the listener to reliably reconstruct the full concept underlying *hand-drawn* despite the omission.

Moreover, the ellipsis construction adds new functionality on its own: it avoids repetition while sharpening the contrast between *hand* and *computer* – highlighting the most salient distinction of the speaker's question. The interplay between ellipsis and compounding here exemplifies constructional integrity in practice: flexible form (even for words), robust meaning.

Now, let's explore another example that shows how constructional flexibility is needed for guiding semantic interpretation in ways that is often missed by word + syntax approaches. Consider example (3), which is typically analyzed in terms of "filler-gap" mechanisms (Sag, 2010):

(3) Do you remember the song that Jack loves?

A standard filler-gap analysis is that the Object *the song* has been "extracted" from its canonical position in an underlying structure like *Jack loves the song*, leaving a silent "gap" behind. But this view presupposes that the sentence is derived from a more basic configuration – and more importantly: it misrepresents the semantics and fails to explain *why* this structure exists in the first place.

What the formal account overlooks is the function of *the song* as a noun phrase. Typically, a definite noun phrase signals that its referent is identifiable (Lambrecht, 1994). If I say *Jack loves the song*, I imply that *the song* alone is sufficient to know which song I am talking about. Yet, in example (3), that is not the case. The only way to identify the song in question is precisely by saying that Jack loves it.

To resolve this, English speakers have invented what we might call a Möbius strip construction: a structure in which the intended interpretation emerges through interconnection rather than strict syntactic hierarchy. Breaking this process down:

1. A definite noun phrase must establish a uniquely identifiable referent. Here, *the song* alone fails to do so. Ironically, this means that the phrase *Jack loves the song* cannot possibly have served as the extraction site: it ticks all checkboxes of syntactic well-formedness, but it is pragmatically incomplete in the current context because its Object NP cannot fulfil its referential function.

2. To restore its functionality, the noun phrase must integrate the transitive clause as post-nominal modifier, effectively "recruiting" it to establish reference (and maintaining its own functional integrity).

3. Because the transitive clause is now embedded within the noun phrase, the object no longer needs to be realized in post-verbal position – it is structurally distributed. This allows the transitive construction to sustain its functional integrity, even as it supports the referential work of the noun phrase.

All of the above illustrates how formal flexibility is not just permitted but sometimes required to maintain functional integrity. Rather than treating structures like (3) as derivations, we should recognize them as adaptations that balance syntactic constraints with communicative needs.

3 Linguistic Form as Dynamic Tensegrity

We just argued that *constructional integrity* is a necessary condition for modelling language as a dynamic, adaptive system. But how can we achieve such integrity in a formal architecture?

We propose **dynamic tensegrity** as a novel architectural metaphor for linguistic form: a system in which constructions interact through interdependent constraints held in dynamic equilibrium.

This section introduces the concept of dynamic tensegrity and explores how it may inform both the theoretical understanding and computational implementation of construction grammar.

3.1 What is Tensegrity?

The term *tensegrity* – short for "tensile integrity" – was first coined by Buckminster Fuller in the mid-20th century to describe an architectural principle in which structural stability arises from the interaction between elements under continuous tension and elements under localized compression (Swanson, 2013). The concept was directly inspired by the sculptural artwork of Kenneth Snelson, a student of Fuller, whose pioneering constructions gave the idea physical form.

Figure 1 illustrates this principle through Snelson's artwork *Tree I*, a suspended structure composed of rigid struts (under compression) and flexible cables (under tension). In a true tensegrity, the struts never touch; they are held in place entirely by the pull of the tensile network. What looks improbable – floating beams in open space – is in fact a precisely tuned equilibrium, where no single element holds the structure together, but the system as a whole sustains its integrity.

Tensegrity exemplifies how complex systems can be resilient without rigidity, and stable without central control – a principle that has found wide applications not only in architecture, but also in robotics (Shah et al., 2022) and biology (Huang et al., 2006; Swanson, 2013).

3.2 Systems within Systems within Systems

The principle of tensegrity has become a powerful heuristic in biomechanics, offering insights across multiple scales of organization – from cells and tissues to organs and whole organisms. This nested hierarchy – "tiers of systems within systems within systems" (Ingber, 2003, p. 1167), with emergent properties at each level – will serve as our architectural model for language.

Figure 1: *Tree I* by Kenneth Snelson. Photo available via Wikimedia Commons, licensed under CC BY-SA 4.0.[1] Image scaled for formatting; no other changes.

The most intuitive application of tensegrity is at the level of the organism. Humans walk, stand, and move their bodies thanks to the musculoskeletal system, which achieves structural stability through the interaction of local compression and continuous tension. In this system, bones bear compression, while muscles, tendons, and ligaments form a tensile network that distributes mechanical stress throughout the body. Crucially, our bones do not directly touch each other for bearing load, but are suspended in a matrix of tension. This distributed arrangement allows the body to absorb shocks, adapt to uneven terrain, and maintain balance. Rather than relying on central control, the system achieves equilibrium through the self-regulating dynamics of its parts. Tensegrity robotics draws directly on these properties, designing systems that are both robust and flexible (Shah et al., 2022).

But tensegrity also applies to other tiers. At the cellular level, for instance, tensegrity may explain how cells retain structural integrity despite constant remodelling in response to external pressure. As Huang et al. (2006) note, the shape and mechanical behavior of mammallian cells are largely governed by an internal scaffold called the "cytoskeleton". In the tensegrity view, this scaffold functions like a dynamic 3D network: filamentous proteins (microfilaments) create tension by pulling inward, while rigid rods (microtubules) push outward to resist compression. Together, these elements form a self-stabilizing system that can deform and reorganize without collapsing.

[1] https://creativecommons.org/licenses/by-sa/4.0/deed.en

Besides observing that tensegrity seems to be a "fundamental design principle that is used to stabilize biological networks at all size scales in the hierarchy of life", Huang et al. (2006, p. 296) argue that tensegrity "may also directly impact information flow in biological systems". In other words, the spatial properties of tensegrities (their *form*) seem strongly interrelated with the *function* of the systems they stabilize.

Form and function co-emerging through distributed tension may also explain how grammatical constructions stabilize meaning. The next step is therefore to see if this structural principle extends from cells to constructions.

3.3 Linguistic Tensegrity Networks

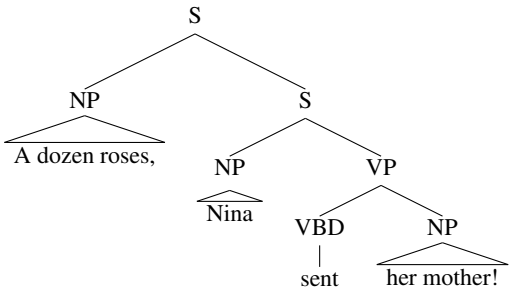
To model linguistic structure as dynamic tensegrities, we must shift from trees to networks: not hierarchical command structures, but lattices of constraint and support.

Network-based thinking already plays a key role in understanding how structure emerges from distributed interaction. In complex adaptive systems, networks model how local interactions give rise to global properties (Yang et al., 2023). In construction grammar, they help chart the usage-based relationships between constructions (Diessel, 2019) and collostructional affinities (Wellens, 2011).

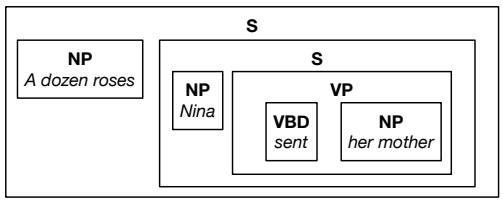
Moreover, networks are suited for describing both *structural* systems and *information processing* systems (Huang et al., 2006). While this paper focuses on structure, tensegrity networks thus offer a promising foundation for modelling the interplay between form and meaning in future research.

Let us start with the familiar tree representation in (4) and its notational variant as boxes-within-boxes in (5).

(4)



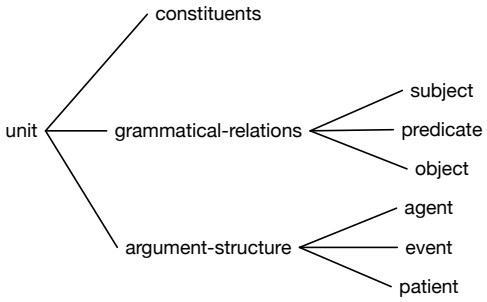
(5)



Although they differ in notation – one emphasizing hierarchy, the other slot-filling – both rest on the same structural assumption that phrase-structural relations form the backbone of linguistic analysis, which can then be enriched with feature-value descriptions. From a mechanical perspective, both representations connect rigid parts directly: structure is assembled by stacking the different building blocks on top of each other. From an information flow perspective, there is a clear entry point to the structure (the root node S); and accessing relevant information requires tree traversal.

By contrast, tensegrity models suggest an alternative: suspension through tension. Structures are not held together by direct contact, but through distributed constraint resolution. Moreover, unlike a tree with a single root and directed paths, a tensegrity-like network behaves more like a city map, offering multiple points of entry and redundant pathways for accessing information.

To formalize this perspective, we reconceptualize the basic components of linguistic structure. Instead of stacking nodes directly on top of each other in a rigid hierarchy, we *suspend* them as separate **compression units**: modular structures that **encapsulate** a coherent set of feature-value pairs. These may describe phonological, morphosyntactic, semantic or pragmatic properties, as illustrated in (6). The unit as a whole behaves as a rigid body: internally structured, but moving or linking like a single entity.



(6)

In this model, constituent structure is not the backbone, but one of many possible descriptions encoded within a compression unit – e.g. using a *constituents* feature. This decoupling allows the model to accommodate the multi-dimensionality of constructions, which often cut across levels in non-uniform ways. Some constructions are compact, engaging with only phonology and morphology. Others reach from the conceptual to the pragmatic. As Fried and Östman (2004, p. 19) put it:

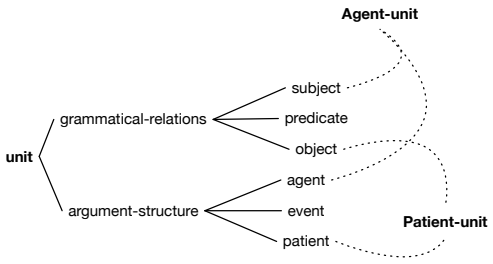
> "Construction Grammar [is] a multi-dimensional framework in which none of the layers is seen as 'more basic' than any other; constructions only differ in the extent to which they make use of these resources."

We now have compression units. But what holds them together? What provides tension?

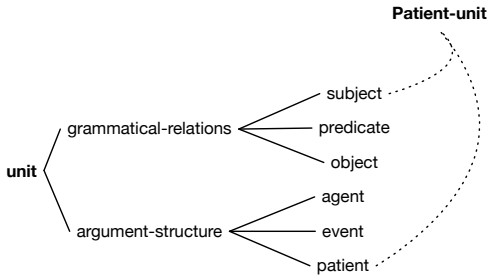
In our model, tension arises through **unit links**. Rather than filling slots directly or imposing hierarchical dominance, compression units are connected by linking the values of their internal features to other compression units. These unit links define the network of **interdependence**, and thus serve as tension constraints: abstract forces that align and balance feature information across units. In sum, structure no longer emerges through assembly, but through **relational suspension**.

Examples (7) and (8) offer a partial illustration of how English active and passive sentences link compression units in different ways. In our model, the Passive is not treated as a transformation of some underlying active form, but as a **self-contained tensegrity configuration** with its own usage and interpretation conditions.

(7)



(8)



Likewise, the topicalization alternation in sentence pair (1) can now be reinterpreted as two distinct tensegrity configurations. In *A dozen flowers, Nina sent her mother*, the same underlying ditransitive relation holds as in *Nina sent her mother a dozen flowers* – but the compression units associated with the topicalized phrase are reoriented through a different pattern of linking, driven by discourse-pragmatic constraints.

Rather than treating such alternations as mere surface permutations, our models captures them as **structurally distinct yet functionally coherent configurations** within a dynamic network. The crucial difference from transformational approaches is that the latter privilege a single base structure from which others are derived, whereas in a tensegrity model, **all constraints coexist on equal footing**. Each configuration emerges from a unique balance of forces, not a uniform derivational origin.

3.4 Constructions vs. Construction Schemas

Formalized approaches to construction grammar, such as Berkeley Construction Grammar (Fillmore, 2013) and Sign-Based Construction Grammar (Michaelis, 2009), describe constructions as static constraints on well-formed tree configurations or filler-slot relations. In our dynamic tensegrity view, we subscribe to the constraint-based approach, but we further adopt a distinction between constructions – the emergent, conventionalized pairings of form and function observable at the community level – and **construction schemas**, the knowledge that an individual language user has about these constructions.

Construction schemas act as **dynamic operators** that build and combine constructions. More specifically, construction schemas need the following components:

- **Applicability conditions**: These determine when a schema may be invoked, typically based on the presence of certain feature-value pairs across one or more compression units.
- **Linking constraints**: Once activated, the construction schema imposes relational constraints that coordinate values across units. These constraints include unit links – the "tensions" that maintain structural integrity.
- **Contributing part**: The schema may also expand existing units or introduce entirely new compression units – adding structure necessary for stabilizing the evolving network and for satisfying the language user's communicative needs.

For example, the schema for building the English Active-Transitive construction requires the presence of three compression units that represent an event and its agent and patient. In the resulting construction, these units are linked together in a dynamic equilibrium.

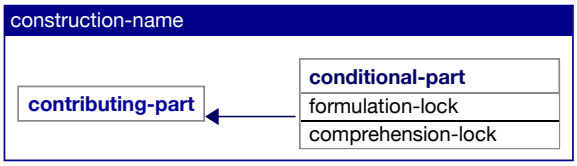


Figure 2: A skeletal representation of a simple construction schema in Fluid Construction Grammar.

4 Modelling Dynamic Tensegrity in Fluid Construction Grammar

Fluid Construction Grammar (FCG) is an open-source computational platform for construction grammar that was originally developed to support experiments in language evolution and emergence (Steels, 2004, 2012). As such, it makes no a priori assumptions about syntactic structure: its architecture is explicitly designed to allow structure to emerge in local usage events.

Another key advantage of FCG is its transparent, symbolic representation of constructions, which are fully inspectable via an interactive web interface (Loetzsch, 2012). Besides its source code, FCG also has a freely available and cross-platform Integrated Development Environment (van Trijp et al., 2022), making it an ideal system for interpretable modelling and hypothesis testing.

Moreover, prior research in FCG has already anticipated several of the architectural intuitions explored in this paper – including the treatment of grammatical structure as dynamic networks rather than rigid trees (Beuls and Steels, 2013; van Trijp, 2016, 2020). The current proposal builds on and extends this work by introducing tensegrity as a unifying metaphor for organizing and linking constructional representations.

4.1 Adopting Tensegrity in FCG

Due to space limitations, this paper focuses on how the tensegrity principle can be adopted within the FCG framework. A working implementation and interactive web demo are available in the supplementary materials.

Figure 2 provides a schematic representation of a construction schema in FCG. Each schema consists of two sides:

- The right-hand side defines the applicability conditions: what must be present in the transient structure to activate the schema.
- The left-hand side provides the contributing part: what information needs to be added.

Each side contains one or multiple boxes – units in FCG parlance – which we reinterpret here as compression units. On the conditional side, each compression unit is further subdivided into two parts: the *formulation-lock* (above) determines the conditions under which the construction can be built in production; while the *comprehension-lock* specifies the necessary cues for recognizing a construction in parsing.

Constructional activation, constraint resolution and contribution are all three achieved through two types of unification processes: matching and merging (Steels and De Beule, 2006). The activation process works by matching the relevant lock's conditions against the **transient structure** – a dynamically evolving representation of the sentence's tensegrity network. If the match is successful, this simultaneously resolves (some) structural tension by unifying variables or completing partial structures. Every successful match is followed by two merging phases: first, the information of the lock is opened and integrated with the transient structure; after which the contributing part is unlocked and integrated as well. The result is an expanded and more stabilized tensegrity network.

4.2 Proof-of-Concept Implementation

In the supplementary materials, the mapping between tensegrity and FCG is illustrated using the following expressions:

(9) a. The rabbit *nibbled* the carrot.
b. The carrot was *nibbled* by a rabbit.
c. A *nibbled* carrot.

Although each example includes the same verb form *nibbled*, traditional analyses consider only (a) as the "basic" verbal form, while (b) and (c) are treated as derivations – respectively, the passive verb form and a deverbal adjective. But this analysis misses important semantic generalizations.

We offer a different interpretation: *nibbled* is never a derived form, but a stable tensegrity network – a local configuration licensed by two constructions: one evoking a semantic frame, the other imposing morpho-aspectual form and semantics.

The NIBBLE-CXN[2] evokes a semantic frame based on the verbal root *nibble*. Following a force-dynamic approach (Talmy, 2000; Croft, 2012), this frame encodes a **causal chain** in which an agent applies force to affect a patient.

[2]The abbreviation "cxn" stands for "construction".

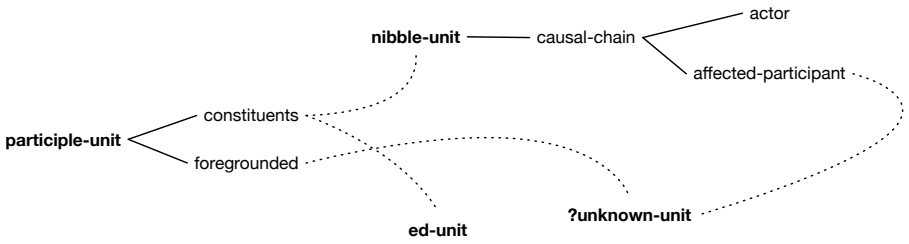
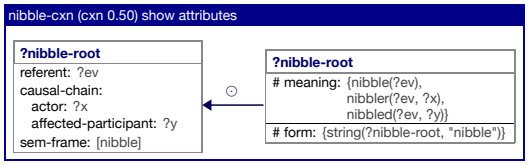


Figure 3: Schematic tensegrity configuration for the verb form "nibbled". Three compression units – representing the participial form, the verb root, and the morphological suffix – are suspended in dynamic equilibrium by unit links. The structure foregrounds the affected participant but remains underspecified (*?unknown-unit*), allowing integration into multiple grammatical configurations (e.g. passive, nominal).

Example (10) offers a simplified representation of the NIBBLE-CXN schema. In plain words, this construction schema will be activated if the speaker wants to express the Nibble Frame (here represented using first-order predicates in the formulation lock); or if the listener observes the verbal root "nibbles" in comprehension. If matching is successful – that is, a compression unit is found in the transient structure that meets the schema's conditions – the contributing information is added. In this case, the schema expands the compression unit with additional information.

(10)



Moving onto the *ed*-suffix, English *ed*-forms consistently foreground the result state of the event – precisely the point in the causal chain where the patient has been affected. This explains why that *ed*-form naturally fits all three examples: (a) in past tense expressions, the event has been completed; (b) in the passive, the focus is on what happened to the undergoer; and (c) in nominal phrases, the *ed*-form identifies the noun as the affected participant.

We therefore model the -ED-PARTICIPLE-CXN as a morpho-aspectual construction: it contributes the surface form "nibbled", and constrains interpretation to highlight the affected participant. In comprehension, its construction schema is activated as soon as a verbal root is encountered followed by the *-ed*-suffix.

The result of combining these two constructions is schematically represented in Figure 3. As can be seen, three compression units are suspended (*participle-unit*, *nibble-unit* and *ed-unit*), held in tension by two unit links going from the participle-unit's constituent feature to the other two units. However, in order to reach full equilibrium, the tensegrity configuration still needs a fourth compression unit (here mentioned using the placeholder variable *?unknown-unit*), indicating that the affected participant is foregrounded by this tensegrity structure. This underspecified unit allows the construction to be integrated in various other tensegrity configurations, such as passive and nominal networks.

This illustrates the central advantage of tensegrity: grammatical structure emerges not from derivational rules, but from locally stable networks of constraints that can be "pulled" or reoriented – robust, flexible, and transparently interpretable.

5 Conclusion

This paper made the case for **constructional integrity** as a necessary condition for modelling language as a complex adaptive system. It then introduced **dynamic tensegrity** as a novel metaphor and modelling principle for ensuring structural integrity in construction grammar.

By shifting from stacked trees to suspended networks, we offered a structural account of grammar that promises both robustness and flexibility without derivations. Our formalization reframes construction schemas as constraint-resolving operators within a tensegrity network of compression units and linking tensions. We supplemented the approach with a proof-of-concept implementation in Fluid Construction Grammar.

The principle of tensegrity complements the constructional modelling landscape by aligning with usage-based, multidimensional views of grammar while remaining human interpretable. Future work may focus on how tensegrity may also stabilize the semantic functions of constructions.

Acknowledgments

The core idea for this paper grew out of discussions about tensegrity robotics with my colleagues David Colliaux, Peter Hanappe and Sébastien Marino; and could not have blossomed without the mentorship of Luc Steels, and the many years of collaboration with the FCG community, particularly Katrien Beuls and Paul Van Eecke. I also thank Peter Hanappe, Hiroaki Kitano, Vittorio Loreto and all my colleagues for creating such a superb working environment.

Supplementary Materials

The demonstration that supports this paper can be downloaded at: https://zenodo.org/records/15778283. The file can be loaded and tested with the FCG Editor, available at https://fcg-net.org.

References

Katrien Beuls and Luc Steels. 2013. Agent-Based Models of Strategies for the Emergence and Evolution of Grammatical Agreement. *PLoS ONE*, 8(3):e58960.

Katrien Beuls and Paul Van Eecke. 2026. Fluid Construction Grammar. In Hilary Nesi and Peter Milin, editors, *International Encyclopedia of Language and Linguistics*, 3d edition edition. In Press.

Ted Briscoe and Ann Copestake. 1999. Lexical rules in constraint-based grammars. *Computational Linguistics*, 25(4):487–526.

Benjamin Bruening. 2018. The lexicalist hypothesis: Both wrong and superfluous. *Language*, 94(1):1–42.

Joan Bybee and Sandra Thompson. 2000. Three Frequency Effects in Syntax. In *Berkeley Linguistics Society*, volume 23, pages 65–85, Dwinelle Hall, Berkeley CA. University of California, Berkeley.

Noam Chomsky. 1970. Remarks on Nominalization. In Roderick A. Jacobs and Peter S. Rosenbaum, editors, *Reading in English Transformational Grammar*, pages 184–221. Ginn, Waltham.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford.

William Croft. 2012. *Verbs: Aspect and Causal Structure*. Oxford University Press, Oxford.

Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge University Press, Cambridge.

Charles J. Fillmore. 1988. The Mechanisms of "Construction Grammar". In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55, Berkeley CA. Berkeley Linguistics Society.

Charles J. Fillmore. 1989. Grammatical construction theory and the familiar dichotomies. In R. Dietrich and C.F. Graumann, editors, *Language Processing in Social Context*, pages 17–38. North-Holland/Elsevier, Amsterdam.

Charles J. Fillmore. 2013. Berkeley Construction Grammar. In *The Oxford Handbook of Construction Grammar*, pages 110–132. Oxford University Press, Oxford.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language*, 64(3):501–538.

Mirjam Fried and Jan-Ola Östman. 2004. Construction Grammar: A Thumbnail Sketch. In Mirjam Fried and Jan-Ola Östman, editors, *Construction Grammar in a Cross-Language Perspective*, pages 11–86. John Benjamins, Amsterdam.

Adele E. Goldberg. 2006. *Constructions At Work: The Nature of Generalization in Language*. Oxford University Press, Oxford.

Adele E. Goldberg. 2019. *Explain Me This. Creativity, Competition, and the Partial Productivity of Constructions*. Princeton University Press, Princeton NJ.

Adele E. Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

Liliane Haegeman. 1994. *Introduction to Government and Binding Theory*, 2nd edition, volume 1 of *Blackwell Textbooks in Linguistics*. Blackwell, Oxford and Cambridge MA.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*, 2nd edition edition. Hodder Education, London.

Sui Huang, Cornel Sultan, and Donald E. Ingber. 2006. Tensegrity, dynamic networks, and complex systems biology: Emergence in structural and information networks within living cells. In *Complex Systems Science in Biomedicine*, pages 283–310. Springer, Boston MA.

Donald E. Ingber. 2003. Tensegrity I: Cell structure and hierarchical systems biology. *Journal of Cell Science*, 116(7):1157–1173.

Knud Lambrecht. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents*. Cambridge Studies in Linguistics 71. Cambridge University Press, Cambridge.

Ronald W. Langacker. 2000. A Dynamic Usage-Based Model. In Michael Barlow and Suzanne Kemmer, editors, *Usage-Based Models of Language*, pages 1–63. Chicago University Press, Chicago.

Stephen Guy Lapointe. 1980. *A Theory of Grammatical Agreement*. Ph.D. thesis, University of Massachusets, Amherst.

Martin Loetzsch. 2012. Tools for Grammar Engineering. In Luc Steels, editor, *Computational Issues in Fluid Construction Grammar*. Springer, Berlin.

Laura A. Michaelis. 2009. Sign-Based Construction Grammar. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 139–158. Oxford University Press, Oxford.

Steven T. Piantadosi. 2024. Modern language models refute Chomsky's approach to language. In *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, pages 353–414. Language Science Press, Berlin.

Ivan A. Sag. 2010. English Filler-Gap Constructions. *Language*, 86(3):486–545.

Dylan S. Shah, Joran W. Booth, Robert L. Baines, Kun Wang, Massimo Vespignani, Kostas Bekris, and Rebecca Kramer-Bottiglio. 2022. Tensegrity robotics. *Soft Robotics*, 9(4):639–656. PMID: 34705572.

Luc Steels. 2004. Constructivist Development of Grounded Construction Grammars. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 9–19, Barcelona. Association for Computational Linguistic Conference.

Luc Steels, editor. 2011. *Design Patterns in Fluid Construction Grammar*, volume 11 of *Constructional Approaches to Language*. John Benjamins, Amsterdam.

Luc Steels, editor. 2012. *Experiments in Cultural Language Evolution*, volume 3 of *Advances in Interaction Studies*. John Benjamins, Amsterdam.

Luc Steels and Joachim De Beule. 2006. Unify and Merge in Fluid Construction Grammar. In P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv, editors, *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Commun*, LNAI 4211, pages 197–223. Springer-Verlag, Berlin.

Randel L. Swanson. 2013. Biotensegrity: A unifying theory of biological architecture with applications to osteopathic practice, education, and research – a review and analysis. *Journal of Osteopathic Medicine*, 113(1):34–52.

Leonard Talmy. 2000. *Toward a Cognitive Semantics, Concept Structuring Systems*, volume 1. MIT Press, Cambridge, Mass.

Remi van Trijp. 2016. Chopping down the syntax tree: what constructions can do instead. *Belgian Journal of Linguistics*, 30(1):15–38.

Remi van Trijp, Katrien Beuls, and Paul Van Eecke. 2022. The FCG Editor: An innovative environment for engineering computational construction grammars. *PLOS ONE*, 17(6):e0269708. Publisher: Public Library of Science.

Remi van Trijp. 2020. Making Good on a Promise: Multidimensional Constructions. *Belgian Journal of Linguistics*, 34:357–370.

Thomas Wasow. 1977. Transformations and the lexicon. In Peter W. Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*, pages 327–360. Academic Press, New York.

Pieter Wellens. 2011. Organizing Constructions in Networks. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*, pages 181–201. John Benjamins, Amsterdam.

Kewei Yang, Jichao Li, Maidi Liu, Tianyang Lei, Xueming Xu, Hongqian Wu, Jiaping Cao, and Gaoxin Qi. 2023. Complex systems and network science: A survey. *Journal of Systems Engineering and Electronics*, 34(3):543–573.

Psycholinguistically motivated Construction-based Tree Adjoining Grammar

Shingo Hattori* **Laura Kallmeyer**** **Rainer Osswald****
*University of Tokyo **Heinrich Heine University Düsseldorf
*shinhattori@g.ecc.u-tokyo.ac.jp
**{kallmeyer,osswald}@phil.hhu.de

Abstract

This paper proposes a formal framework based on Tree Adjoining Grammar (TAG) that aims to incorporate central tenets of Construction Grammar while integrating mechanisms from a psycholinguistically motivated variant of TAG. Central ideas are (i) to give TAG-inspired tree representation to various constructions including schematic constructions like argument structure constructions, (ii) to link schematic constructions that are extensions of each other within a network of constructions, (iii) to make the derivation proceed incrementally, (iv) to allow the prediction of upcoming constructions during derivation and (v) to introduce the *incremental extension* of schematic constructions to larger ones via *extension trees* in a usage-based manner. The final point is the major novel contribution, which can be conceptualized as the on-the-fly traversal of the inheritance links in the network of constructions. Moreover, we present first experiments towards a parser implementation. We report preliminary results of extracting constructions from the Penn Treebank and automatically identifying constructions to be added during incremental parsing, based on a generative language model (GPT-2).

1 Introduction

Theories of construction grammar (Goldberg, 1995, 2003) posit that the building blocks of language are constructions, or form-meaning pairs at various levels of abstraction: not only words but also phrasal or larger patterns, from multi-word expressions and collocations to syntactic patterns like argument structures. In this approach, those constructions are combined to form representations linked to sentences in a manner constrained by semantic or pragmatic compatibilities as well as usage. It is further hypothesized that these constructions are cognitively organized as a network, whose links represent inheritance relations.

Despite the strong concern with cognitive plausibility in construction grammar, psycholinguistic evaluation of its tenets seems to be done mainly on qualitative predictions (Bencini and Goldberg, 2000; Perek, 2025), while more quantitative evaluation with psycholinguistic data has been attempted for other grammar formalisms (Roark et al., 2009; Padó, 2007; Konieczny, 1996; Stanojević et al., 2023; Brennan et al., 2016). This is not surprising, given that the existing formalized variants of construction grammar (Bergen and Chang, 2005; Steels, 2017; Boas and Sag, 2012) and studies of computational extraction of constructions (Dunn, 2017) appear to lack broad coverage and psycholinguistically plausible parsers.

In this regard, Tree Adjoining Grammar (TAG, Joshi et al., 1975) seems to be a promising framework to formalize and implement construction grammar, as has been suggested in Kallmeyer and Osswald (2013) and Lichte and Kallmeyer (2017) among others. Moreover, there is a psycholinguistically motivated variant of TAG with an incremental broad-coverage parser (Psycholinguistically motivated TAG, PLTAG, Demberg et al., 2013; Demberg-Winterfors, 2010). PLTAG, however, does not take into account all the key tenets of construction grammar.

Thus, we will develop Psycholinguistically motivated Construction-based TAG, or PLCxTAG, a formalization of construction grammar inspired by PLTAG. In addition, we will present a preliminary implementation of the framework leveraging a neural language model (LM) as a proof of concept, including a lexicon automatically extracted from the Penn Treebank (Marcus et al., 1993, PTB) and a broad-coverage supertagger based on a decoder LM (GPT-2, Radford et al., 2019), which will comprise the core of an incremental parser for psycholinguistic evaluation to be conducted in future work.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 130–143

2 Related work

2.1 Construction grammar

Key tenets of construction grammar. The focus of our approach is on the following five key tenets of construction grammar (Goldberg, 2003). First, the grammar is viewed as the *composition of constructions*, which are form-meaning pairs stored in memory. This requires that phrasal or larger patterns can be directly associated with meaning. Also, the constructions are combined depending on the semantic compatibility among them, rejecting the autonomy of syntax. Second, *schematic constructions* such as argument structure constructions (Goldberg, 1995) are also memorized. These capture the regularities traditionally described in syntax. Third, construction grammar is generally *surface-oriented* (Goldberg, 2003). Great emphasis is placed on the surface generalization, without resorting to assumptions about deep structure from which surface structures might be derived. Also, phonologically null elements like traces, PRO or null function heads are avoided. Fourth, the theories of construction grammar often take a *usage-based approach* (Langacker, 1987; Bybee, 2010; Tomasello, 2005), which postulates that specific usages are memorized according to their frequencies, and more general constructions like schematic constructions arise in a bottom-up manner. Finally, the lexicon of constructions is postulated to be structured as a *network of constructions* (Diessel, 2023), where constructions are connected based on inheritance relations among them, where more specific constructions "inherit" information from abstract constructions. This network captures how schematic constructions, such as argument structure constructions, productively license quite rare but grammatical uses, e.g. "sneeze the foam off the cappuccino" or "kick Bob a ball" (Goldberg, 1995).

Formalizations of construction grammar. There are three major computational theories of construction grammar: Embodied Construction Grammar (ECG, Bergen and Chang, 2005; Chang, 2008; Feldman, 2022; Bryant, 2008), Fluid Construction Grammar (FCG, Steels, 2017; Beuls and Van Eecke, 2023) and Sign-Based Construction Grammar (SBCG, Boas and Sag, 2012). All of them have parser implementations, but no incremental parser exists for FCG nor SBCG to our knowledge, though Müller (2017) suggests that existing incremental parsers for Head-driven Phrase Structure Grammar (e.g. Konieczny, 1996) could be adapted to SBCG. ECG does have a psycholinguistically motivated incremental parser, "constructional analyzer" (Bryant, 2008), but the scalability of the parser appears limited due to the need of manually writing the grammar and defining parameters for some phenomena of interest (Bryant, 2008, p. 156).

There have also been attempts to extract constructions automatically with a view to making the study of constructions scalable and not limited to a handful of constructions selected by linguists (Dunn, 2017). Still, it is by no means obvious how these constructions can be combined to form actual sentence representations.

2.2 Modeling human sentence processing

Properties of human sentence processing. Accumulating studies in psycholinguistics have not only identified various psycholinguistic phenomena, such as garden path, indicating the preferences of certain structures over others, but also demonstrated three general principles of human sentence processing (Demberg and Keller, 2019). First, the parse is built *incrementally*, updated for every new word (Konieczny, 2000; Tanenhaus et al., 1995). Second, it is known that the mismatch of subject and reflexive pronoun affects the reading time even before the VP is completed with the second PP object, suggestive of *connected syntactic structure* facilitating such agreement (Sturt and Lombardo, 2005). Finally, parsing proceeds based on *predictions*, e.g. by anticipating the argument of a verb before encountering it (DeLong et al., 2005; Kamide et al., 2003; Staub and Clifton, 2006).

PLTAG. PLTAG is a psycholinguistically motivated variant of TAG (Demberg et al., 2013; Demberg-Winterfors, 2010), which is designed to satisfy the three properties described above.

There are crucial innovations to maintain incrementality and connectedness during derivation, such as the prediction-verification scheme. Moreover, the grammar has been automatically extracted from the PTB, and a broad-coverage parser was implemented based on it, which was then evaluated on reading time data.

Yet, there is some room for exploring alternative formalizations, and more importantly, PLTAG does not satisfy some key tenets of construction grammar, e.g., there is no network of constructions

and null elements are used extensively (though this latter property is not inherent to the PLTAG formalism, i.e., it would be straightforward to choose a grammar without null elements).

3 Formal framework

Our formalization is guided by the principles of linguistic and psycholinguistic plausibility. Conditions for *linguistic plausibility* consist of the five tenets of construction grammar: (i) Grammar as the composition of constructions, (ii) Schematic constructions, (iii) Surface-oriented approach, (iv) Usage-based approach and (v) Network of constructions. As conditions for *psycholinguistic plausibility*, three properties of human sentence processing are chosen: (a) incremental, (b) connected and (c) predictive.

Our formalization incorporates on the one hand aspects of *constructions and their composition*, and on the other hand aspects of incremental processing that lead to *incremental extension of constructions* (along the network of constructions) and additional *prediction of upcoming constructions*.

CxTAG: Constructions and their composition. Our starting point is the use of (lexicalized) TAG (LTAG) and frame semantics for modeling constructions along the lines of Kallmeyer and Osswald (2013) and Lichte and Kallmeyer (2017). In that approach, the elementary trees of TAG are paired with frame-semantic representations (formalized as extended attribute value structures) to *elementary constructions* (or *lexicalized constructions*), in which specific nodes of the tree can be linked to specific components of the semantic frame. The tree-combining operations *substitution* and *adjunction* go along with the unification of the associated frames. (In our case, *substitution* and *sister adjunction* are used.[1]) In the present paper, we keep the semantic side of constructions largely aside since our primary focus is on the formal aspects of incremental syntactic processing as well as on the extraction of the form aspect of constructions from treebanks. Note, however, that in the ongoing parsing implementation (Section 5), semantics is implicitly covered both at the lexical as well as at the constructional level via the embeddings learned in the model.

[1] *Substitution* consists of inserting a tree at a non-terminal leaf, i.e., filling an argument slot. *Sister adjunction* merges the root of the adjoining tree with an internal node, thereby introducing additional subtrees below that internal node.

Elementary trees are partial constituent trees such that each tree has at least one leaf representing the head word, called a *lexical anchor*, and that all of the anchor's projections and arguments are localized in the same tree where arguments are represented as leaves that are substitution nodes; see the tree for 'gave' in Fig. 1 for illustration.[2]

Our theoretical judgment of what qualifies as arguments or adjuncts is more or less in line with standard LTAG analysis (XTAG Research Group, 1998): The subject and the objects of verbs and the noun phrase in prepositional phrases are arguments, while determiners, adjectives, adverbs, auxiliary verbs, semi-auxiliary verbs (e.g. "used to"), copula verbs, raising verbs, complementizers and the infinitive marker "to" are adjuncts.

Elementary constructions also cover multiword expressions, collocations, and frequently co-occurring patterns, motivated by usage-based postulates. In these cases, the corresponding elementary trees can have multiple anchors, called *co-anchors*.

Schematic constructions such as argument structure constructions, however, are not to be represented by LTAG elementary trees since they are unlexicalized and lack a lexical anchor. Therefore, in order to represent them, we employ unlexicalized counterparts of elementary trees known as *supertags* in the TAG literature (Bangalore and Joshi, 2010). The parent node of a removed lexical element in a supertag is called an *anchor node* and is usually marked with a $\diamond$.

The network of constructions and the 'extend' operation. Within the (L)TAG framework, more complex constructions can be derived from simpler ones in a strictly compositional manner by general tree operations such as substitution and adjunction. How different elementary constructions are related to each other and, in particular, how certain constructions can be part of certain other constructions, are not expressed by tree operations but at a different level of grammatical description, often called the *metagrammar* (Kallmeyer and Osswald, 2013; Lichte and Kallmeyer, 2017). The metagrammar allows the specification of trees (and frames) by means of expressive constraint languages (Crabbé et al., 2013; Lichte and Petitjean, 2015).

The resulting set of schematic constructions can

[2] The specific categories and trees used in this paper are to some degree influenced by the PTB (Marcus et al., 1993) employed in Section 5. Notice, however, that the formalization presented here is general and compatible with other constituency formats.

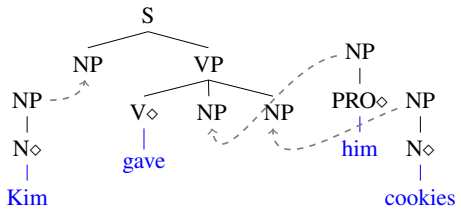


Figure 1: Composition of elementary trees for (1-a); the black trees represent schematic constructions (i.e., 'supertags'); dashed arrows represent tree operations.

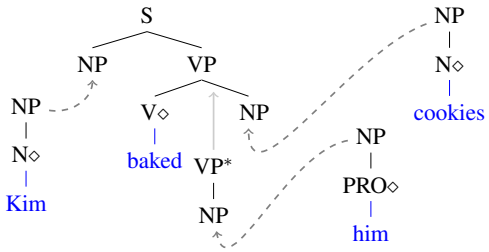


Figure 2: Elementary trees and schematic constructions for (1-b); the solid gray arrow indicates extension by a benefactive NP resulting in the ditransitive construction.

be seen as a *network of constructions*, whose relations indicate specialization ("inheritance") but also more complex types of embeddings of substructures. The described division of labor between general operations on elementary constructions and the more advanced "off-line" specification of elementary constructions and their interrelation by means of constraints have conceptual and practical advantages. The approach falls short, however, if we are to study in which way the network of constructions can guide incremental language processing.

In order to overcome this problem, we propose an additional "operation" *extend*, which mimics standard tree operations, mostly adjunction, but in effect realizes the move from one construction to another, usually more extended construction, which (at least on the semantic side) is typically non-compositional. We refer to the modified formalism as *Construction-based Tree Adjoining Grammar (CxTAG)*.

Schematic, i.e., lexically unanchored constructions are instantiated by lexicalized constructions for words in context. Therefore, instead of treating elementary constructions as part of the lexicon, we further assume here that for a given lexical element w_i, depending on its left context LC_i (comprising $w_1 \ldots w_i$ and any syntactic, semantic and pragmatic structures built so far), schematic constructions $t_i^{\diamond}$ are chosen with a certain probability $P(t_i^{\diamond}|LC_i)$, and also extensions $t_{i,j}^{e}$ of previously chosen constructions $t_j^{\diamond}$ $(j < i)$ to more specific ones occur with a certain probability $P(t_{i,j}^{e}|LC_i)$.[3] This is why trees assigned to each word in the figures contain anchor nodes with $\diamond$.

Consider the examples in (1) for illustration, whose derivations are shown in Figs. 1 and 2.

(1) a. Kim gave him cookies
b. Kim baked him cookies

Both sentences are assumed to give rise to the same syntactic trees except for the lexical head verb. Their derivations differ, however: We may assume that the verb 'gave' generally selects a ditransitive argument structure construction with much higher probability than a transitive construction. The ditransitive construction then provides substitution sites for the two remaining NP arguments. The verb 'baked', by comparison, would select a transitive argument structure construction with higher probability by default as is most likely, and a benefactive NP tree is added by means of extend, which in turn gives rise to a structure that matches an existing construction, namely the benefactive ditransitive construction. In this case, we call the added NP construction (the benefactive NP) an *extension tree* and say that the transitive construction has been extended to the benefactive ditransitive construction.

Note that in general extension trees could also add another co-anchor to extend constructions to those representing multi-word expressions. From the perspective of usage-based approach, extension trees can be seen as secondary generalizations that emerge through the comparison of existing schematic constructions, which are themselves generalizations from instantiations, along with the formation of the network, and they are often interpretable as constructions themselves.[4]

PLCxTAG: Incremental extension of constructions and prediction of upcoming constructions. In the following, we extend the CxTAG formalism in the spirit of PLTAG (Demberg et al., 2013). So far, the order of the derivation steps in CxTAG is

[3] In our implementation, the probabilities for schematic constructions are estimated via fine-tuning a GPT-2 model towards predicting them, see Section 5.

[4] We avoided referring to extension trees as constructions categorically, since extension trees may not always qualify as independent constructions from a semantic viewpoint.

not restricted, but in order to achieve psycholinguistic plausibility, we extend the formalism towards allowing derivations that build connected parses incrementally. This not only imposes constraints on how syntactic operations can be applied but also requires additional mechanisms and operations.

At the same time, our formalism departs from PLTAG in that it aims to capture the key tenets of construction grammar. Crucially, the extension of schematic constructions is an integral part of the incremental derivation: For each word, a supertag representing some schematic construction is added given the context up to it, and it can be extended later to match the appropriate construction by the end of the sentence in a way described in CxTAG. This *incremental extension* might be conceptualized as the traversal of inheritance links during the derivation.

The overall idea of our psycholinguistically motivated modification of CxTAG is that at each word, we add a new elementary tree and at most one extension tree via substitution or sister adjunction, where the operation can be in both directions (i.e., the already derived tree added to the new one by substitution/adjunction or vice versa). These derivation steps are restricted in such a way as to add material only to the right of the rightmost lexical node in the already derived tree. As an example, Fig. 3 shows the sequence of derived trees we obtain with such a derivation when combining the constructions from Fig. 2. The (orange) tree fragment representing the supertag added at 'baked' is extended to a larger supertag at 'him'. The words (in green) above the $\rightsquigarrow$ arrows indicate the next word, whose processing triggers the next derivation step.

Such an incremental connected derivation is, however, not always possible: When the elementary tree of a word should be combined with a node in an elementary tree of a future word, it is impossible to create a connected partial parse. For example, in "John often smiles", the supertags for words up to "often" cannot be combined without the S and VP nodes from the supertag for "smiles" (see Fig. 4).

To remedy this, we employ a restricted version of the prediction-verification scheme proposed in PLTAG (Demberg et al., 2013; Demberg-Winterfors, 2010) and a new scheme, delay, to compensate for the restriction.

The prediction-verification scheme consists of two steps: *prediction* and later *verification*. In prediction, for each word, an additional supertag is optionally selected as a *prediction tree*, such that it contains all the nodes missing at that point but necessary for a connected partial parse. We assume here that the prediction trees are chosen probabilistically, depending on the left context.[5] They are attached via substitution or sister adjunction to the partial parse while keeping track of the fact that they are only predicted. At a later stage, such a prediction tree has to be verified by a matching supertag that is anchored by an actual word. Note that a supertag used to verify a prediction tree will not be attached to the partial parse via substitution or adjunction. Instead, the nodes from the prediction tree have to be mapped to corresponding nodes in the verifying supertag in such a way that labeling and structural relations are preserved. A sample derivation is given in Fig. 5. The prediction tree is the upper-left tree (depicted in gray) and the mapping performed in the verification operation is indicated by dotted arrows. The red numbers at the arrows indicate the order of derivation operations.

Prediction trees can be extended to larger su-

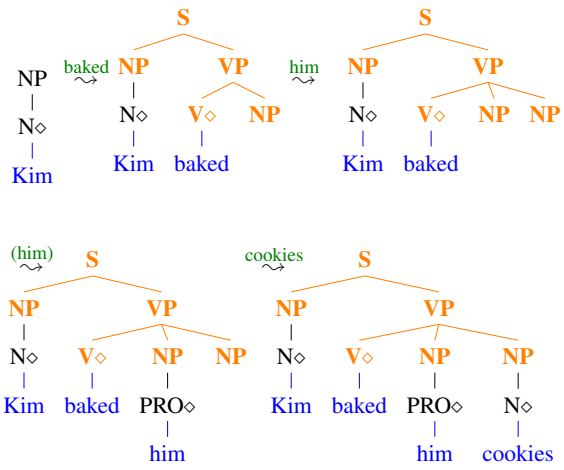


Figure 3: Incremental and connected derivation for (1-b): Incremental extension for inheritance

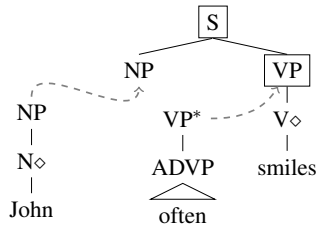


Figure 4: Intervening nodes come from the supertag for a subsequent word

[5]Estimated by the second classification head of the fine-tuned GPT-2 model in our implementation, see Section 5.

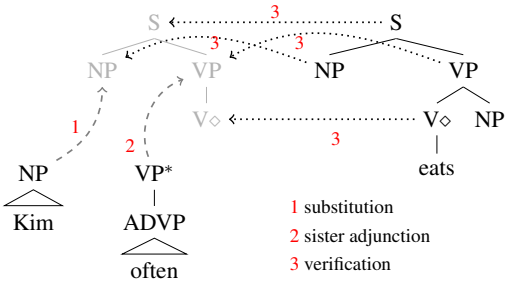


Figure 5: PLCxTAG derivation of 'Kim often eats ...'

pertags due to verification, in cases where the two trees are not isomorphic. For instance, to attach an adverb after a subject NP, as in Fig. 5, it is enough to use an intransitive supertag as the prediction tree, even if the following verb is actually transitive. In this case, the prediction tree is extended during verification due to the verb's transitive supertag.

For prediction-verification, it is an open question how to configure the granularity of predictions: Even though we decided to use supertags as prediction trees, one could also create a separate lexicon of tree fragments that only contain the necessary nodes, as in PLTAG (Demberg et al., 2013; Demberg-Winterfors, 2010).

On the other hand, we forbid adding several prediction trees in a row, following PLTAG. This means that if the nodes needed for a connected partial parse come from multiple supertags, one prediction tree is not enough in our framework. In the example in Fig. 6, the boxed AP and NP nodes are both necessary in order to combine the supertag for 'very' with the partial parse.

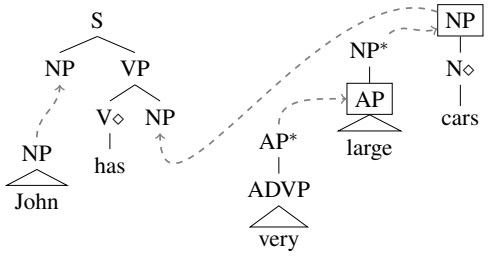


Figure 6: Nodes from multiple supertags

To address such cases, we decided to relax the incrementality condition and allow the delayed attachment of a word's supertag: The creation of a connected partial parse is suspended, waiting for necessary nodes to appear in supertags assigned to subsequent words. To be more precise, we attach the supertag in question to the supertag for the next word first, which in turn will be combined with the partial parse. If needed, we might allow further delays. Our hypothesis is that most actually occurring cases are covered with a maximal delay of 1, based on the inspection of the data from the PTB used in Section 5 below.

The resulting extension of CxTAG is called *Psycholinguistically motivated CxTAG (PLCxTAG)*.

4 Sample derivations involving various constructions

For further illustrations, let us look at a few more interesting examples. It should be noted that the derivations presented below are not prescriptive, and PLCxTAG can be employed to represent alternative analyses.

Argument structure constructions without co-anchors. Let us first consider examples of argument structure constructions: caused motion construction and resultative construction.

(2) a. Kim kicked the ball over the fence
b. Kim sneezed the foam off the cappuccino
c. Kim painted the barn red
d. Kim kicked his feet sore

Derivations for (2-a) and (2-b) are given in Fig. 7–8. The red numbers indicate the order of derivation steps. Dashed arrows indicate substitutions and sister adjunctions that are standard combinations of elementary trees. In contrast, solid gray arrows indicate operations that extend an already chosen elementary tree to a larger one such as the caused motion construction or the resultative construction with extension trees. For the sake of readability, some of the sub-derivations are omitted, i.e., only their result is displayed.

In Fig. 7, the transitive tree selected for 'kick' is extended to the caused motion construction by adding a path PP.

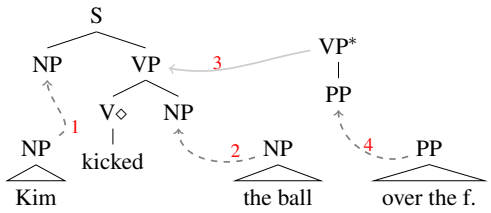


Figure 7: PLCxTAG derivation of (2-a)

The derivation of (2-b) extends an intransitive tree (the 'sneezed' supertag) to the caused motion

construction where slots for both mover (NP) and path (PP) are added. The derivation for resulta-

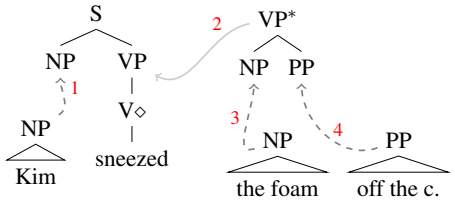


Figure 8: PLCxTAG derivation of (2-b)

tive constructions such as (2-c) and (2-d) would look very similar to the two derivations in Fig. 7 (where the object NP is already introduced with the verb) and Fig. 8 (where the object NP is introduced via the extension), except that the result is an AP. Semantically, caused motion and resultative constructions differ as a matter of course.

Constructions with co-anchors. In the following, we will discuss analysis options for two examples involving co-anchors, (3-a) and (3-b).

(3) a. Kim elbowed his way through the crowd
b. Kim kicked the bucket

The respective complete elementary trees for the two verbal construction would be as in Fig. 9. Note, however, that (3-a) is not restricted to a single verb while 'kick the bucket' is a fixed idiomatic expression. Concerning the latter, it can also have a literal meaning, but our analysis here is about the idiomatic meaning, to which we assign an independent elementary tree with co-anchors.

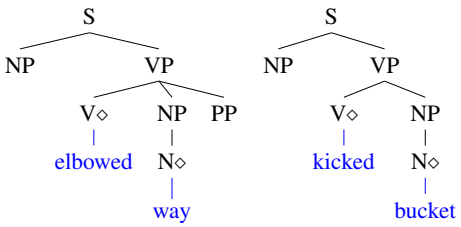


Figure 9: Complete elementary trees for the multi-anchored constructions in (3-a) and (3-b)

If we assume a strictly incremental derivation with prediction trees whenever words cannot be connected yet, we could choose an analysis as in Fig. 10. Step 5 in this case is special since it not only verifies the predicted NP tree but also reanalyzes its substitution (operation 3 in this derivation) as a substitution that is an extension. This latter is something that is not yet covered by the above definition of PLCxTAG but that could be added.

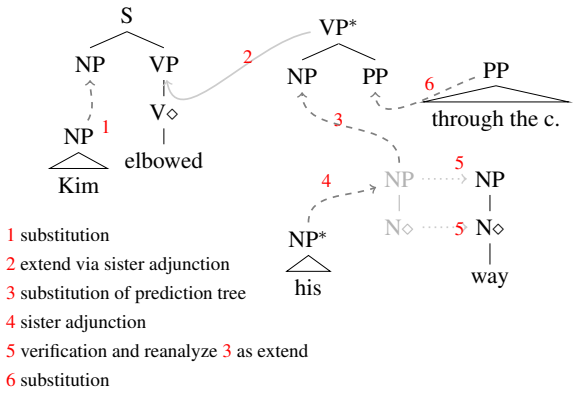


Figure 10: PLCxTAG derivation of (3-a) with verifcation and reanalyze as extension

The difficulty here comes from the fact that 'his' has to be attached before seeing 'way', a difficulty that could be avoided with a delay for this attachment. In general, it might be justified to adopt a delay for all cases of functional operator attachment. If we do this, we can actually adopt an analysis as in Fig. 11, where the extension tree anchored at 'way' extends the verbal tree.

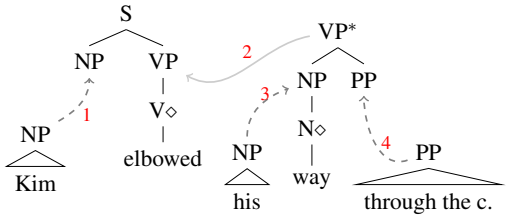


Figure 11: PLCxTAG derivation of (3-a) with delayed attachment of 'his'

Similarly, for (3-b), we could predict an NP tree at 'the', followed by a verification by a tree anchored by 'bucket', thereby reanalyzing the substitution of the prediction tree as an extension. Assuming, however, that the attachment of function words can be delayed, this complication is not needed. The corresponding derivation is given in Fig. 12.

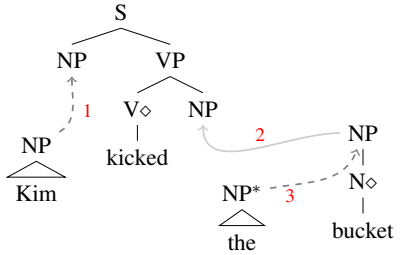


Figure 12: PLCxTAG derivation of (3-b) with delayed attachment for 'the'

5 Implementation

In this section, we will present preliminary results from the ongoing implementation of a PLCxTAG parser. The details of the parser architecture are still in development, but the core components are to be the lexicon, the (k-best) supertagger and a parallel parsing scheme with beam search. As the current stage of the implementation, we present the preliminary lexicon extraction and a supertagger.

Lexicon extraction. To extract the lexicon automatically, we used the Sections 00–24 from the Wall Street Journal portion of the Penn Treebank (Taylor et al., 2003). The extraction procedure is similar to those previously used for LTAG extraction (Xia et al., 2000; Chiang, 2000; Demberg et al., 2013): Trees in the PTB were preprocessed to suit our linguistic analysis and nodes were marked as head, argument and adjunct, using a modified version of the head and argument/adjunct rules from Collins (1999, 1997).

The preprocessing of the trees consists of five steps, three of which were conducted before marking the nodes, and the remaining two were performed afterwards.

Firstly, to be surface-oriented, (a) we deleted all null elements (Bies et al., 1995), including traces and PRO. Secondly, (b) we collapsed unary branches that appear due to the previous step, while retaining those which are already present in the original PTB. Thirdly, (c) we relabeled the part of speech tags of auxiliaries 'have', 'be' and 'do' as AUX, which are originally labeled as full verb, because all auxiliaries including those should be labeled as adjunct.

Then, we annotated each node of the trees according to the head/argument/adjunct rules described in Appendix A, making use of the tags including function tags.

At this point, (d) we reduced the tagset by removing function tags and merging some of the tags used in the PTB (cf. Appendix B). This reduced the number of tags from 71 to 36, which would in turn reduce the number of supertags and thus potentially improve the efficiency of the supertagger training as well as the performance of the resulting model. Then, (e) we collapsed the tree branches if the label of the parent node is identical to that of the head child node and no other children nodes are labeled as argument. This is because in CxTAG sister adjunction is used to attach adjuncts directly to the head phrase, without introducing new branches.

Fig. 13 illustrates the procedure with an example from the PTB. First, (a) the * (PRO) under -NONE- is removed, and then (b) the resulting unary branch from S to VP is collapsed. According to the marking rules, all nodes (except the root) are labeled as H(ead), C(omplement for argument) and A(djunct). Finally, (d) the tagset is reduced, where function tags like -SBJ are removed, NNP is modified along with other subcategories of noun to N(oun) and VB and VBD are merged into V(erb). Then, (e) the VP above TO and its head child labeled as VP are collapsed, since the other child is an adjunct. The result is the middle tree in Fig. 13.

Then elementary trees were extracted based on the annotation in a bottom-up fashion. Basically, the tree is to be split at the nodes labeled as C or containing children labeled as A (cf. the third tree in Fig. 13).

The elementary trees in this version are without co-anchors, excluding some well-known constructions like way-construction. Also, extension trees and hence the network of constructions are not covered yet. For those, we would need further extraction procedures to combine or decompose supertags obtained so far, depending on the statistics of the entire treebank.

In addition to supertags, we also extracted a sequence of prediction trees for each sentence in the data that the parser has to predict when processing it by computing the connection path (Demberg et al., 2013; Demberg-Winterfors, 2010) to check for each pair of adjacent words if there are some intervening nodes belonging to supertags to be anchored by subsequent words. At this point, the delay mechanism is not implemented yet, limiting the coverage to 33466 sentences out of 49208.

Our current lexicon extraction on the PTB yields 2663 different supertags, out of which 1293 are also used as prediction trees.

Supertagging. The supertagger is implemented via fine-tuning GPT-2 using multi-task learning with two classification heads, returning a pair of prediction tree (possibly none) and supertag for each word. We trained the model on Sections 02–21 for five epochs and evaluated it on Section 23. The data consists of a sequence of pairs of prediction tree/none and supertag. For more details see Appendix C.

The per-word accuracy of the supertagger after training is 0.91 and 0.79 for prediction trees

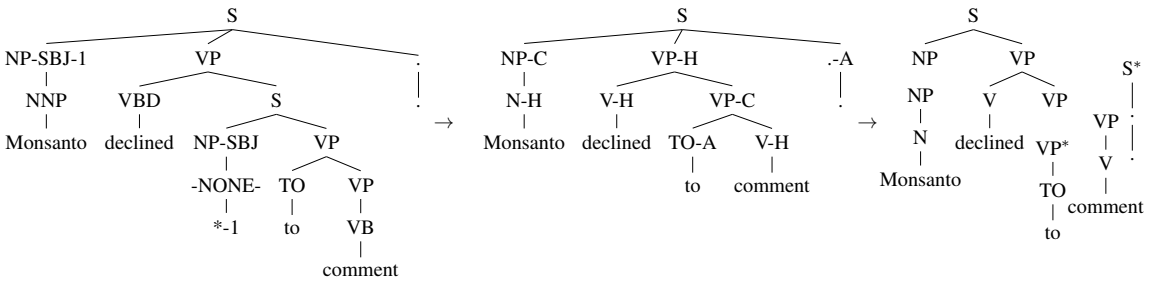


Figure 13: Sample extraction of supertags from a PTB tree

and supertags, respectively. Note that the accuracy for prediction trees looks better than it actually is, since for most words, the prediction is none, i.e., always predicting none could already yield an accuracy of about 0.80. For supertags, our accuracy is below the state-of-the-art for LTAG supertagging (for instance, Bladier et al., 2019, achieved 0.81 on French, which is usually harder than English), but not comparable to standard supertagging results because (for the sake of psycholinguistic plausibility) we employed a generative incremental LM as basis while Bladier et al. (2019) used a bidirectional model. Overall, the scores we achieved with these first experiments are quite promising.

6 Discussion and Conclusion

Summary. In this paper, we presented an alternative formalization of construction grammar guided by linguistic and psycholinguistic plausibility.

In particular, incremental extension based on frequencies of constructions is a novel way to formalize the inheritance and underspecification under usage-based tenets: The selection of schematic constructions is distributed over multiple words, facilitated by the the network of constructions and reflecting the predictability of constructions at each point in the incremental derivation.

Also, the results of our preliminary implementation of PLCxTAG, extracted lexicon and the supertagger, serve as a proof of concept. We are hopeful that future implementation of PLCxTAG will pave the way for quantitative psycholinguistic evaluation of the tenets of construction grammar.

Future directions. We are currently building a PLCxTAG parser which we will use to quantify the processing difficulty in a way similar to (Demberg et al., 2013; Demberg-Winterfors, 2010) for comparison with reading time data.

To this end, we are in the process of modifying the extraction and supertagging implementation to include extension trees and a delay mechanism, as well as designing the parallel parsing scheme that decides how to combine the trees returned by the supertagger. Concerning extension trees, the idea is to start with the supertags extracted in the way proposed here and train our supertagger on it. Based on the predicted supertags, we will then decompose some of the extracted gold supertags into smaller supertags and extension trees.

For psycholinguistic evaluation, we plan to use a corpus annotated with reading time data such as that presented in Frank et al. (2013) and evaluate along the lines of Mielczarek et al. (2025).

In addition, there are some aspects of the formalism that might require further discussion and improvement. For instance, we have yet to see which of the strategies sketched in Section 5 works better for constructions with co-anchors. In this context, the evaluation on psycholinguistic data will be taken into consideration. Also, there are some syntactic phenomena beyond the current formalization. For example, the use of only substitution and sister adjunction restricts the generative capacity of PLCxTAG in such a way that phenomena of long-distance dependencies cannot be adequatly treated. Second, we did not explicitly model semantic representation of constructions. Note, however, that our supertagger produces semantic representations of lexical anchors and, implicitly in its activation vectors, also of schematic constructions.

Finally, we are planning to extend PLCxTAG to other languages, in particular German and French, where we already have experience with TAG-based supertag extraction (Bladier et al., 2019). Ideally, in the long run, we would like to apply the framework also to a typologically broader set of languages such as Japanese.

Acknowledgements

We would like to thank three anonymous reviewers for their valuable and helpful feedback.

References

Srinivas Bangalore and Aravind K. Joshi, editors. 2010. *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*. MIT Press, Cambridge, MA.

Giulia M L Bencini and Adele E Goldberg. 2000. The contribution of argument structure constructions to sentence meaning. *J. Mem. Lang.*, 43(4):640–651.

Benjamin K Bergen and Nancy Chang. 2005. Embodied construction grammar in simulation-based language understanding. In *Constructional Approaches to Language*, pages 147–190. John Benjamins Publishing Company, Amsterdam.

Katrien Beuls and Paul Van Eecke. 2023. Fluid construction grammar: State of the art and future outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 41–50, Washington, D.C. Association for Computational Linguistics.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for treebank II style Penn treebank project. Technical report, Linguistic Data Consortium.

Tatiana Bladier, Jakub Waszczuk, Laura Kallmeyer, and Jörg Janke. 2019. From partial neural graph-based LTAG parsing towards full parsing. *Computational Linguistics in the Netherlands Journal*, 9:3–26.

Hans C. Boas and Ivan Sag, editors. 2012. *Sign-Based Construction Grammar*. CSLI Publications, Stanford.

Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain Lang.*, 157-158:81–94.

John E. Bryant. 2008. *Best-Fit Constructional Analysis*. Ph.D. thesis, University of California at Berkeley.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press, Cambridge.

Nancy Chang. 2008. *Constructing grammar: A computational model of the emergence of early constructions*. Ph.D. thesis, University of California at Berkeley.

David Chiang. 2000. Statistical parsing with an automatically-extracted Tree Adjoining Grammar. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 456–463, Hong Kong. Association for Computational Linguistics.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Madrid, Spain. Association for Computational Linguistics.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Processing*. Ph.D. thesis, University of Pennsylvania.

Benoit Crabbé, Denys Duchier, Claire Gardent, Joseph Le Roux, and Yannick Parmentier. 2013. XMG: eXtensible MetaGrammar. *Computational Linguistics*, 39(3):591–629.

Katherine A DeLong, T Urbach, and M Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.*, 8:1117–1121.

Vera Demberg and Frank Keller. 2019. Cognitive models of syntax and sentence processing. In Peter Hagoort, editor, *Human Language: From Genes and Brains to Behavior*, pages 293–312. The MIT Press.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated Tree-Adjoining Grammar. *Computational Linguistics*, 39(4):1025–1066.

Vera Demberg-Winterfors. 2010. *A Broad-Coverage Modelof Prediction in Human Sentence Processing*. Ph.D. thesis, University of Edinburgh.

Holger Diessel. 2023. *The Constructicon: Taxonomies and Networks*. Elements in Construction Grammar. Cambridge University Press, Cambridge.

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

Jerome A Feldman. 2022. Advances in embodied construction grammar. In *Benjamins Current Topics*, pages 147–167. John Benjamins Publishing Company, Amsterdam.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabrielle Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

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 Cogn. Sci.*, 7(5):219–224.

Aravind K Joshi, Leon S Levy, and Masako Takahashi. 1975. Tree adjunct grammars. *J. Comput. Syst. Sci.*, 10(1):136–163.

Laura Kallmeyer and Rainer Osswald. 2013. Syntax-driven semantic frame composition in Lexicalized Tree Adjoining Grammars. *Journal of Language Modelling*, 1(2):267–330.

Yuki Kamide, Gerry T M Altmann, and Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *J. Mem. Lang.*, 49(1):133–156.

Lars Konieczny. 1996. *Human sentence processing: a semantics-oriented parsing approach*. Ph.D. thesis, University of Freiburg.

Lars Konieczny. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites*. Stanford University Press, Stanford, CA.

Timm Lichte and Laura Kallmeyer. 2017. Tree-Adjoining Grammar: A tree-based constructionist grammar framework for natural language understanding. In *The AAAI 2017 Spring Symposium on computational construction grammar and natural language understanding*, pages 205–212, Stanford, CA.

Timm Lichte and Simon Petitjean. 2015. Implementing semantic frames as typed feature structures with XMG. *Journal of Language Modelling*, 3(1):185–228.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330. Special Issue on Using Large Corpora: II.

Lukas Mielczarek, Timothée Bernard, Laura Kallmeyer, Katharina Spalek, and Benoit Crabbé. 2025. Modelling expectation-based and memory-based predictors of human reading times with syntax-guided attention. In *Proceedings of BriGAP-2*, Düsseldorf, Germany. Association for Computational Linguistics.

Stefan Müller. 2017. Head-driven phrase structure grammar, sign-based construction grammar, and fluid construction grammar: Commonalities and differences. *Constructions and Frames*, 9(1):139–173.

Ulrike Padó. 2007. *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Ph.D. thesis, Saarland University.

Florent Perek. 2025. Behavioral evidence and experimental methods. In Mirjam Fried and Kiki Nikiforidou, editors, *The Cambridge Handbook of Construction Grammar*, Cambridge Handbooks in Language and Linguistics, page 196–219. Cambridge University Press.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Miloš Stanojević, Jonathan R Brennan, Donald Dunagan, Mark Steedman, and John T Hale. 2023. Modeling structure-building in the brain with CCG parsing and large language models. *Cogn. Sci.*, 47(7):e13312.

Adrian Staub and Charles Clifton, Jr. 2006. Syntactic prediction in language comprehension: evidence from either...or. *J. Exp. Psychol. Learn. Mem. Cogn.*, 32(2):425–436.

Luc Steels. 2017. Basics of fluid construction grammar. *Constructions and Frames*, 9(2):178–225.

Patrick Sturt and Vincenzo Lombardo. 2005. Processing coordinated structures: incrementality and connectedness. *Cogn. Sci.*, 29(2):291–305.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeillé, editor, *Treebanks. Building and Using Parsed Corpora*, pages 5–22. Springer, New York.

Michael Tomasello. 2005. *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M Rush. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv [cs.CL]*.

Fei Xia, Martha Palmer, and Aravind Joshi. 2000. A uniform method of grammar extraction and its applications. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* -, Morristown, NJ, USA. Association for Computational Linguistics.

XTAG Research Group. 1998. A Lexicalized Tree Adjoining Grammar for English. Technical report, University of Pennsylvania, Institute for Research in Cognitive Science.

A Appendix: Annotation of nodes as head, argument and adjunct

As is the case in previous attempts to extract LTAG from the PTB, we exploited the original PTB tags including function tags to mark nodes of trees as head, argument and adjunct.

A.1 Head rules

For the identification of heads, we followed the general procedure described in Collins (1999), where two head percolation tables are used, one for most tags and another for NP.

Still, we have modified both tables (cf. Tables 1 and 2). In particular, three major changes were made to the table for most tags.

Firstly, MD, TO and IN have been removed from the head candidates, since modal auxiliaries, the infinitive marker "to" and complementizers (labeled as IN along with prepositions) are to be adjuncts.

Secondly, -PRD is added as the candidate, since it indicates the existence of accompanying copulative verb like 'be' or 'seem'. In those cases, these verbs are to be adjuncts, even though they are labeled as full verb. That is why -PRD is placed higher in priority than tags for full verbs.

Thirdly, WHNP, WHPP, WHADVP, WHADJP and DT are removed from the candidate list for SBAR.

A.2 Argument/adjunct rules

After annotating the heads, we marked the remaining nodes as either argument or adjunct. Our rules for arguments and adjuncts are inspired by Collins (1997), but there are important changes to the original procedure.

Collins (1997) marks only the following as argument, while marking all else as adjunct:

(a) 1. NP/SBAR/S under S
2. NP/SBAR/S/VP under VP
3. S under SBAR
if without any of the following function tags: -ADV, -VOC, -BNF, -DIR, -EXT, -LOC, -MNR, -TMP, -CLR and -PRP

(b) the first child following the head under PP

This procedure, however, is highly problematic for numerous cases of coordination (especially when no CC or CONJP is involved) and for PP nodes with three or more children, as is exemplified in Fig. 14. In the left-hand side example, two

Parent	From	Priority list
ADJP	L	NNS QP NN $ ADVP JJ VBN VBG ADJP JJR NP JJS DT FW RBR RBS SBAR RB
ADVP	R	RB RBR RBS FW ADVP TO CD JJR JJ IN NP JJS NN
CONJP	R	CC RB IN
FRAG	R	
INTJ	L	
LST	R	LS :
NAC	L	NN NNS NNP NNPS NP NAC EX $ CD QP PRP VBG JJ JJS JJR ADJP FW
PP	R	IN TO VBG VBN RP FW
PRN	L	
PRT	R	RP
QP	L	$ IN NNS NN JJ RB DT CD QP JJR JJS
RRC	R	-PRD VP NP ADVP ADJP PP
S	L	-PRD VP S SBAR ADJP UCP NP
SBAR	L	S SQ SINV SBAR FRAG
SBARQ	L	SQ S SINV SBARQ FRAG
SINV	L	-PRD VBZ VBD VBP VB VP S SINV ADJP NP
SQ	L	-PRD VBZ VBD VBP VB VP SQ
UCP	R	
VP	L	-PRD VBD VBN VBZ VB VBG VBP VP ADJP NN NNS NP
WHADJP	L	CC WRB JJ ADJP
WHADVP	R	CC WRB
WHNP	L	WDT WP WP$ WHADJP WHPP WHNP
WHPP	R	IN TO FW

Table 1: Head table for most phrasal tags, the 2nd column gives the search order (starting from L(eft) or R(ight))

From	Candidate list
R	NN
L	NNP NNPS
R	NNS NX JJR PRP
L	NP
R	$ ADJP PRN
R	CD
R	JJ JJS RB QP

Table 2: Head table for parent tag NP

S children are coordinated by a semicolon labeled as :, where the first S is already labeled as head due to the head rule described in Table 1. In this case, the latter S would be marked as argument of the former, which it is not. The second example shows an instance where the annotator placed D and N immediately below PP without intermediate NP, resulting in D being marked as argument and N as adjunct.

Figure 14: Problematic examples from the PTB.

Therefore, for (a), when candidate nodes are coordinated, we only choose the left-most one, and for (b), we decided to use finer-grained conditions, depending on the number of PP's children.

In addition to these, we decided to mark all the nodes with some function tags like -SBJ as argument.

The resulting rules are:

(a) 1. NP/SBAR/S under S
2. NP/SBAR/S/VP under VP
3. S under SBAR

if without any of the following function tags: -ADV, -VOC, -BNF, -DIR, -EXT, -LOC, -MNR, -TMP, -CLR and -PRP

&

i. if not coordinated

or

ii. if the left-most coordinated element

(b) 1. the non-head child under a PP with two children
2. the first NP child under a PP with three or more children

(c) nodes with one of the following function tags: -DTV, -BNF, -LGS, -PUT, -SBJ, -CLF and -CLR

B Appendix: Tagset reduction

Tagset reduction was done by collapsing tags according to Tables 3 and 4.

Original tags	Reduced tag
JJ JJR JJS	A
RB RBR RBS WRB	Adv
DT PDT WDT PRP$ WP$	D
CD NN NNS NNP NNPS PRP WP EX $ #	N
AUX MD VB VBP VBZ VBN VBD VBG	V
Other POS tags	(unchanged)

Table 3: Tagset reduction for POS tags

Original tags	Reduced tag
ADJP WHADJP	AP
ADVP WHADVP	ADVP
NP NAC NX QP WHNP	NP
PP WHPP	PP
S SQ SBAR SBARQ SINV	S
Other phrasal tags	(unchanged)

Table 4: Tagset reduction for phrasal tags

C Appendix: Training of the supertagger

C.1 Model architecture

We modified GPT2PreTrainedModel (Radford et al., 2019) from the transformers library (Wolf et al., 2019) by adding the second linear classification head. The overall loss function was the mean of two cross entropy functions, one for each classifier.

C.2 Hyperparameters used in the training

We used the Trainer class from transformers library to train the model. See Table 5 for values chosen for the hyperparameters used in the training.

hyperparameter	value
learning rate	2e-05
number of epochs	5
weight decay	0.01
train batch size	8
evaluation batch size	8
seed	42
betas for ADAMW	(0.9,0.999)
epsilon for ADAMW	1e-08

Table 5: Hyperparameters of supertagging

Assessing Minimal Pairs of Chinese Verb-Resultative Complement Constructions: Insights from Language Models

Xinyao Huang
ECNU Shanghai &
HHU Düsseldorf
huangxinyao_23
@stu.ecnu.edu.cn

Yue Pan
ECNU Shanghai
51270400014
@stu.ecnu.edu.cn

Stefan Hartmann
HHU Düsseldorf
hartmast@hhu.de

Yanning Yang
ECNU Shanghai
ynyang@english
.ecnu.edu.cn

Abstract

Chinese verb-resultative complement construction (VRCC), constitute a distinctive syntactic-semantic pattern in Chinese that integrates agent-patient dynamics with real-world state changes; yet widely used benchmarks such as CLiMP and ZhoBLiMP provide few minimal-pair probes tailored to these constructions. We introduce ZhVrcMP, a 1,204 pair dataset spanning two paradigms: resultative complement presence versus absence, and verb–complement order. The examples are drawn from *Modern Chinese* and are annotated for linguistic validity. Using mean log probability scoring, we evaluate Zh-Pythia models (14M-1.4B) and Mistral-7B-Instruct-v0.3. Larger Zh-Pythia models perform strongly, especially on the order paradigm, reaching 89.87% accuracy. Mistral-7B-Instruct-v0.3 shows lower perplexity yet overall weaker accuracy, underscoring the remaining difficulty of modeling constructional semantics in Chinese.

1 Introduction

Chinese verb-resultative complement constructions (VRCC) stand out as one of the distinctive and challenging features in syntax and semantics. They feature a complex interplay of elements like agent-patient dynamics, resultative states, and real-world state changes. Any syntactic or semantic mismatch in these constructions can sharply reduce sentence acceptability (often marked with *), as it diminishes the likelihood of such events occurring in reality. For illustration, example (a) shows a clear relation between agent and patient. The agent "I (我)" performs the action "broke (打)" on the patient "vase", which causes the state change "up (碎)" and yields a complete resultative event. The physical properties of the patient constrain the result: a vase can plausibly become "broke up (碎)" but not "into two pieces (断)", so (b) is well formed in syntax but infelicitous in meaning. VRCC also respect event order, causing action must come first and the result must follow, so (c) violates this sequence and is semantically unacceptable. Capturing VRCC requires balancing the individual semantics of components with their overall integration, which poses significant hurdles for grammatical annotation, semantic parsing, and broader NLP applications.

a. 我打碎了花瓶。
wǒ dǎ suì le huāpíng
I broke up the vase.

b. * 我打断了花瓶。
wǒ dǎ duàn le huāpíng
I broke the vase into two pieces.

c. * 我碎打了花瓶。
wǒ suì dǎ le huāpíng
I up broke the vase.

Beyond computational capacity and data scale, the capability of language models to handle complex grammatical structures significantly

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 144–150

impacts their performance in 'understanding' and generating natural language. The minimal pair (MP) method, a foundational linguistic paradigm for testing human language aptitude, has been widely adopted to evaluate language models (LMs) (Xiang et al., 2021; Song et al., 2022; Someya and Oseki, 2023; Warstadt et al., 2023; Capone et al., 2024; Liu et al., 2024). This method generates sentence pairs differing in a single grammatical feature (e.g., word order, morphology, syntax) to assess model comprehension of specific grammatical phenomena. An effective LM should assign higher acceptability probabilities to grammatically and semantically valid sentences in MPs.

With advantages in rigorous variable control, scalable automated design, cross-lingual applicability, and prompt-interference immunity, MPs-based benchmarks exist for multiple languages, including Chinese-specific CLiMP (Xiang et al., 2021), SLING (Song et al., 2022), and ZhoBLiMP (Liu et al., 2024).

Although these datasets excel in broad syntactic paradigm coverage, they lack in-depth exploration of linguistic phenomena through a constructional lens as well as semantic minimal pair design. Poor differentiation between formal and functional competencies leaves model comprehension of semantic relations unaddressed (Mahowald et al., 2024), weakening evaluation interpretability.

CLiMP covers five VRCC paradigms (51000 pairs) but relies solely on complement alteration, with non-random sampling and limited variation compromising validity. In contrast, SLING addresses 38 linguistic phenomena, but omits explicit VRCC. ZhoBLiMP includes partial VRCC in its 14 verb phrase paradigms (14×300 pairs) but lacks a dedicated design and has severely restricted lexis.

To fill this gap, we present ZhVrcMP, a MP dataset for Chinese VRCC, comprising two paradigms and 1,204 total MPs. Words in ZhVrcMP are linguistically selected from the *Modern Chinese*, with partial lexicon adaptation from ZhoBLiMP (Section 3). We tested two types of language models on ZhoBLiMP, our benchmark for assessing how well these models handle Chinese grammar through pairs of sentences with only one grammatical or semantical difference. The first is Zh-Pythia, a set of models adapted for Chinese based on the Pythia framework (Liu et al., 2024), with sizes ranging from 14 million to 1.4 billion parameters (a measure of each model's complexity and capacity). The second is Mistral-7B-Instruct-v0.3, a leading model that has been specially adjusted to follow user instructions effectively; it uses a Transformer design with a 32,768 vocabulary. (Section 4).

Results are detailed in Section 5, along with the part of ZhVrcMP, and model evaluation scripts.

2 Related Work

2.1 Verb-Resultative Complement Construction in Chinese

VRCC, a major subtype of Chinese verb–complement patterns, has the form V + RC where the complement encodes the resultant state caused by the event. This tight coupling of lexical semantics and causation makes VRCC an informative minimal-pair testbed for evaluating LMs' syntactic–semantic processing(Marvin and Linzen, 2018; Kuribayashi et al., 2024).

Construction grammar (CxG)'s form-meaning pairing principle guides the design of minimal pairs (MPs) to probe language models' (LMs) capabilities in semantic role labeling and constructional structure recognition (Weissweiler et al., 2023). As reviewed in recent computational syntheses (Doumen et al., 2024), these principles are operated through unsupervised learning methods (e.g., word embedding clustering) for automatic VRCC identification and

association-based algorithms (e.g., the ΔP metric) for selecting representative MPs (Dunn, 2022). By association-based methods we mean corpus measures that quantify the strength of pairing between verbs and resultative complements, such as ΔP, PMI, and log-likelihood (Stefanowitsch and Gries, 2003; Dunn, 2024). In this paper we construct ZhVrcMP via controlled grammatical manipulations and manual validation rather than corpus-based association scores, although the two approaches are complementary.

Cognitive studies show that VRCC processing involves real-time structure-meaning mapping, with type-shifting complements prolonging model inference (Xue et al., 2021). Thus, VRCC's markedness (e.g., grammaticality constraints) and semantic subtypes (e.g., resultative/stative) enable controlled MPs to assess LMs' grammaticality judgment and low-frequency construction learning (Someya and Oseki, 2023; Warstadt et al., 2023).

2.2 Construction Grammar in Evaluation of Language Capabilities of LMs

CxG grounds LM research through its form-meaning pairing principle, which in turn allows for addressing traditional models' failure to capture implicit constructional information in Chinese VRCC (Zhan, 2017). For instance, Weissweiler (2023) demonstrates that Transformer self-attention aligns with CxG's gestalt cognition, thereby enabling more effective encoding of constructional knowledge and ultimately improving recognition of Chinese VRCC.

These CxG-inspired LM approaches (e.g., Tseng (2022)'s 17.6% accuracy boost in structured tasks) enhance low-frequency construction learning. Despite their importance for LM assessment, the acquisition of constructional knowledge still lacks standardized benchmarks. Existing models focus on form-meaning pattern extraction (Dunn, 2023) and verb argument structure learning (Dominey, 2005; Alishahi and Stevenson, 2008), which form the basis for MP design in VRCC evaluation.

3 Data

ZhVrcMP includes two paradigms: resultative complement presence/absence (Para 1) and verb-complement order inversion (Para 2) with 602 MPs each and 1,204 in total. Curated from the authoritative grammar book *Modern Chinese* (Huang and Li, 2012), which provides comprehensive explanations and examples of Chinese syntax, it adapts the lexicon from ZhoBLiMP. Linguists annotated noun/verb/complement features, generated matching lists (3.1). MPs were automatically generated using an algorithm and manually validated afterwards (3.2).

3.1 Minimal Pairs Generation

3.1.1 Data Sources

As mentioned above, ZhVrcMP sources two main datasets:

1. examples from *Modern Chinese* (pp.78–83);

2. the lexicon of ZhoBLiMP.

For *Modern Chinese* sentences, we parsed components into nouns, verbs, and resultative complements, systematically identifying all verb-complement pairings to ensure dataset richness in capturing VRCC syntactic-semantic relationships.

3.1.2 Vocabulary

ZhVrcMP's noun lexicon has 342 entries with POS, subcategory, gender, animacy, and number annotations. The verb set includes 53 verbs annotated for compatible resultative complements, subject/object subtypes, transitivity, and animacy constraints, matched with 66 unique complements. Using a "subject + verb + complement + (aspect

marker 了) + object" structure, a Python script generated 24,000+ MPs. To minimize unnecessary variation in subjects (which would increase generation workload without adding evaluative value), , "张三 (Zhang San)" was fixed as the sole subject for consistent evaluation.

3.2 Manual Validation

Two annotators with a background in linguistics conducted a double-blind verification of lexical annotations and the automatically generated MPs, yielding an initial inter-annotator agreement rate of 62.6%. After revising 99 pairs (adding indirect objects, aspect markers, etc.) and re-verifying, 602 sentence pairs were selected for each category with a 98% agreement rate (Table 1). We binary-labeled all sentences as GOOD if they were grammatically and semantically well-formed, or as BAD if they differed minimally from the GOOD sentences in one aspect, making them grammatically or semantically invalid. Chi-square tests confirmed statistical equivalence of auxiliary features across label groups (all $p > 0.05$), demonstrating no significant association between exogenous feature distributions and VRCC to isolate the core test variable (Table 2).

4 Models and Methods

We evaluated Zh-Pythia (14M-1.4B) and Mistral-7B-Instruct-v0.3 (4.1) using mean log probability to compare GOOD/BAD sentence probabilities (4.2).

4.1 Models

We evaluated two models: Zh-Pythia (from the ZhoBLiMP study) and Mistral-7B-Instruct-v0.3. Zh-Pythia consists of 20 Chinese-focused Transformer models, trained from scratch on 3B tokens with GPT-NeoX architecture and a Chinese tokenizer to analyze scaling effects on Chinese linguistic phenomena in ZhoBLiMP. Mistral is a commercial 7B-parameter English model optimized for instruction tasks, pre-trained without Chinese adaptation (Table 3).

The models were selected for their contrasting attributes: Zh-Pythia is a Chinese-specific, scalable design evaluated on ZhoBLiMP, while Mistral features a fixed-scale, English-oriented architecture.

4.2 Evaluation Methods

To evaluate the model, we devised a score based on the mean log probability P_{ML}.

$$P_{ML} = \frac{\log P_m(\gamma)}{n_\gamma} \tag{1}$$

In (1), P_{ML} is the mean log probability, $logP_m(\gamma)$ is the mean log probability of model m for γ, n_γ is the sentence count in γ.

Based on the mean log probability obtained above, for each pair set p, we calculated the evaluation score via (2).

$$S(p) = \frac{1}{|p|} \sum_{g,u \in p} \mathbf{1}_{[0,+\infty)}\left(\log \frac{P_{ML}(g)}{P_{ML}(u)}\right) \tag{2}$$

In (2), $S(p)$ is the score for pair set p, $|p|$ is the pair count, and g, u represent the GOOD and BAD sentences, respectively. An indicator function counts the number of valid ratios where $P_{\mathrm{ML}}(g)/P_{\mathrm{ML}}(u) > 1$ (i.e., the model assigns higher probability to the GOOD sentence), and this count is then averaged across pairs to measure the model's ability to capture linguistic capabilities.

Finally, we computed perplexity via mean log probability to quantify model prediction uncertainty, where lower values indicate better data fit.

$$P_{PL} = e^{-P_{ML}} \tag{3}$$

P_{ML} is the mean log probability obtained above. $e^{-P_{ML}}$ converts the log probability to perplexity, a standard metric for assessing LM performance.

	Para 1	Para 2
Number	602	602
GOOD	张三摔破额头。	张三搞错观点。
	Zhāng Sān shuāi pò é tóu	Zhāng Sān gǎo cuò guāndiǎn
	Zhang San fell and broke his forehead.	Zhang San got the wrong point.
BAD	* 张三摔额头。	* 张三错搞观点。
	Zhāng Sān shuāi é tóu	Zhāng Sān cuò gǎo guāndiǎn
	Zhang San fell his forehead.	Zhang San wrong got the point.

Table 1: ZhVrcMP paradigms and example minimal pairs. Para 1 tests resultative complement presence versus absence; Para 2 tests verb–complement order. Each paradigm contains 602 minimal pairs (1,204 total). GOOD is grammatical and semantically plausible; BAD differs only in the targeted constructional feature and is unacceptable (marked with *).

Feature	χ^2	p-value	Conclusion
AM	0.013	0.910	Indep
CL	0.001	0.972	Indep
IO	0.066	0.797	Indep
PASS	0.066	0.797	Indep
MOD	0.639	0.424	Indep

Table 2: Chi-Square Test for Auxiliary features in ZhVrcMP. AM: Aspect Marker, CL: Classifer, IO: Indirect Object, PASS: Passive Voice, MOD: Modifier. Independence is abbreviated as Indep.

Models	Zh-Pythia	Mist
Parameters	14M, 70M, 160M, 410M, 1.4B	7B

Table 3: Evaluation LMs. Mist indicates Mistral-7B-Instruct-v0.3.

5 Results

Results for Zh-Pythia and Mistral-7B-Instruct-v0.3 in the two paradigms are presented in Tables 4, 5, 6. In general, the correct evaluation counts of the models cluster between 400-570 pairs, with scores ranging from 60 to 95, showing a relatively wide range. The perplexity sheds light on the uncertainty of the models in VRCC processing. Although overall scores are high, indicating a notable uncertainty in distinguishing VRCC, substantial differences in perplexity between GOOD and BAD sentences allow effective differentiation.

5.1 Paradigm Results

In Table 4, we analyze performance as a function of model size (number of parameters). In Para 1, Zh-Pythia shows a positive parameter-scaling trend: the number of correctly judged pairs and the mean score both rise with increasing model size. In Para 2, we observe no clear parameter-scaling effect; performance fluctuates slightly across sizes. Overall, Para 2 outperforms Para 1, suggesting that verb–complement order inversion is easier for these models than resultative-complement presence/absence. Performance peaks at 160M (574 correct pairs; mean score 95.19) and declines for larger models, indicating non-monotonic (inverse) scaling beyond 160M. We hypothesize this may relate to training budget or regularization rather than an inherent property of Transformers; order sensitivity can often be captured by local attention patterns, whereas presence/absence relies more on lexical compatibility.

	Zh-Pythia					Mist	Human
Parameter	14M	70M	160M	410M	1.4B	7B	
Para 1	414	487	500	499	517	382	590
Para 2	522	558	574	560	566	504	590
Overall	468	522.5	537	529.5	541.5	443	590

Table 4: Correct Evaluation Numbers of all LMs and human on ZhVrcMP. Human indicates linguistics experts annotation results.

	Zh-Pythia					Mist	Human
	14M	70M	160M	410M	1.4B	7B	
Para 1	68.77	80.90	83.06	82.89	85.88	63.46	98.00
Para 2	86.57	92.54	95.19	92.87	93.86	83.58	98.00
Overall	77.67	86.72	89.13	87.88	89.87	73.52	98.00

Table 5: Percentage Score of all LMs and human on ZhVrcMP

Parameter	**Para 1**		**Para 2**		**Overall**	
	GOOD	BAD	GOOD	BAD	GOOD	BAD
14M	1931.78	2866.34	1928.92	4583.57	1930.35	3724.96
70M	1512.07	2690.21	1511.60	4389.68	1511.84	3539.95
160M	1277.36	2080.77	1278.03	4246.92	1277.70	3163.85
410M	1468.29	2759.91	1469.66	4339.49	1468.98	3549.7
1.4B	2115.22	3902.21	2115.75	6672.02	2115.49	5287.12
7B*	520.36	760.97	524.48	753.70	522.42	757.34

Table 6: Perplexity of all LMs on ZhVrcMP. 7B indicates Mistral-7B-Instruct-v0.3.

5.2 Model Results

Zh-Pythia demonstrates superior performance with smaller parameter sizes compared to Mistral-7B-Instruct-v0.3 (Table 5). Despite Mistral's larger parameters, its correct evaluation counts and scores trail behind Zh-Pythia, particularly in Para 1. However, Mistral exhibits significantly lower perplexity values than Zh-Pythia across both paradigms (Table 6). This duality suggests that while Zh-Pythia's parameter——scaling efficiency aligns more closely with VRCC, Mistral's larger model capacity enhances confidence in distinguishing grammatical and ungrammatical sentences, as reflected by its lower perplexity. The contrasting trends in accuracy and perplexity underscore the interplay between training data relevance and model architectural inductive biases, where Mistral's Transformer design excels in capturing sequence dependencies, thereby reducing perplexity.

6 Conclusion

This paper has introduced ZhVrcMP, a Chinese verb-resultative construction dataset, to assess LMs' semantic-grammatical comprehension. Comprising 1,204 minimal pairs across two paradigms, we have evaluated Zh-Pythia (14M–

1.4B) and Mistral-7B-Instruct-v0.3. The models excel more at verb-complement order than resultative complement presence/absence. Zh-Pythia shows parameter-performance correlation in the latter, peaking at 160M for the former. Mistral lags behind Zh-Pythia, especially in resultative complement tasks. Both models trail human performance, highlighting that construction-semantic processing still has room to improve.

Acknowledgments

This research was funded in part by East China Normal University, 2025 Program for Outstanding Doctoral Student Academic Innovation (Grant No. YBNLTS2025-049).

References

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive science*, 32(5):789–834.

Luca Capone, Alice Suozzi, Gianluca E Lebani, Alessandro Lenci, et al. 2024. BaBIEs: A benchmark for the linguistic evaluation of italian baby language models. In *Ceur Workshop Proceedings*. CEUR-WS.

Peter Ford Dominey. 2005. From sensorimotor sequence to grammatical construction: Evidence from simulation and neurophysiology. *Adaptive Behavior*, 13(4):347–361.

Jonas Doumen, Veronica Juliana Schmalz, Katrien Beuls, and Paul Van Eecke. 2024. The computational learning of construction grammars: State of the art and prospective roadmap. *Constructions and Frames*.

Jonathan Dunn. 2022. Cognitive linguistics meets computational linguistics: Construction grammar, dialectology, and linguistic diversity. *Data Analytics in Cognitive Linguistics: Methods and Insights*, 41:273.

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

Jonathan Dunn. 2024. *Computational construction grammar: A usage-based approach*. Cambridge University Press.

Borong Huang and Wei Li. 2012. *Modern Chinese (Volumn 2)*. BEIJING BOOK CO. INC.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. *arXiv preprint arXiv:2402.12363*.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. ZhoBLiMP: A Systematic Assessment of Language Models with Linguistic Minimal Pairs in Chinese.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. Sling: Sino linguistic evaluation of large language models.

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

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. Cxlm: A construction and context-aware language model. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6361–6369.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023. BLiMP: The Benchmark of Linguistic Minimal Pairs for English.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2023. Explaining pretrained language models' understanding of linguistic structures using construction grammar. *Frontiers in Artificial Intelligence*, 6:1225791.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A Benchmark for Chinese Language Model Evaluation.

Wenting Xue, Meichun Liu, and Stephen Politzer-Ahles. 2021. Processing of complement coercion with aspectual verbs in mandarin chinese: Evidence from a self-paced reading study. *Frontiers in psychology*, 12:643571.

Weidong Zhan. 2017. On theoretical issues in building a knowledge database of chinese constructions. *J. Chinese Inform. Process*, 31:230–238.

Meaning-infused grammar: Gradient Acceptability Shapes the Geometric Representations of Constructions in LLMs

Supantho Rakshit
Dept of ECE / Princeton University
r.supantho@princeton.edu

Adele E. Goldberg
Dept of Psychology / Princeton University
adele@princeton.edu

Abstract

The usage-based constructionist (UCx) approach to language posits that language comprises a network of learned form-meaning pairings (*constructions*) whose use is largely determined by their meanings or functions, requiring them to be graded and probabilistic. This study investigates whether the internal representations in Large Language Models (LLMs) reflect the proposed function-infused gradience. We analyze representations of the English Double Object (DO) and Prepositional Object (PO) constructions in Pythia-1.4B, using a dataset of 5000 sentence pairs systematically varied by human-rated preference strength for DO or PO. Geometric analyses show that the separability between the two constructions' representations, as measured by energy distance or Jensen-Shannon divergence, is systematically modulated by gradient preference strength, which depends on lexical and functional properties of sentences. That is, more prototypical exemplars of each construction occupy more distinct regions in activation space, compared to sentences that could have equally well have occured in either construction. These results provide evidence that LLMs learn rich, meaning-infused, graded representations of constructions and offer support for geometric measures for representations in LLMs.

1 Introduction

A central tenet of usage-based constructionist (UCx) approaches is that our knowledge of language consists of a structured inventory of constructions — conventionalized pairings of form *and* function at varying levels of complexity and abstraction (Goldberg, 2006). The framework posits that language is learned from experience, with contexts and frequencies of use shaping dynamic "ConstructionNets" (Goldberg, 2024) in the minds of speakers. Grammaticality is not a binary state but a continuum of acceptability, an observation supported by work in experimental and computational work on language (Francis, 2022; Gibson and Fedorenko, 2013; Hu et al., 2024).

Here we focus two English constructions, the Double Object (DO) construction (e.g., *She gave the boy the book*) and the Prepositional Object (PO) alternative (*She gave the book to the boy*). We build on a long-standing and widespread focus in linguistics on the combination of information structure and lexical factors that speakers use to choose between the DO and PO constructions: (e.g., Bresnan et al., 2007; Goldberg, 1995; Green, 1974; Levin, 1993; Oehrle, 1976; Wasow and Arnold, 2003). In particular, the recipient argument in the DO strongly tends to be already under discussion and expressed by a definite word (often a pronoun) or short phrase; the transferred entity in a DO, on the other hand, is within the focus domain and is more often expressed by an indefinite noun phrase, which can be a longer phrase. These information structure properties partially emerge from the fact that the DO construction is used to convey real or metaphorical transfer to an animate entity, and animate entities are more likely to be topical in discourse (people often talk about people), while the transfered entity is more likely to be in the focus domain (for a degree of dialect variation see Bresnan and Nikitina, 2009).

The PO construction has been argued to be a subcase of a much broader "caused-motion" construction (Goldberg, 1995, 2002) that can convey a change of location as well as transfer of possession (e.g., *She kicked the ball to him/the wall*). This idea is supported by recent computational work offering a tool for analyzing word meanings in different contexts (Ranganathan et al.,

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 151–157

2025) using interpretable semantic features (Chronis et al., 2023). Ranganathan et al. (2025) report that the features associated with word embeddings vary systematically, depending on whether a given word appears in the DO or PO. In particular, features related to personhood are stronger when the same word, e.g., *London* is the recipient of the DO (*e.g., She sent London the painting*), while features related to location are stronger when *London* appears in the PO (e.g., *She sent the painting to London*).

Verbs' lexical biases also play a role in whether people prefer the DO or PO. The verb *give* is more common in the DO construction, and in fact *give* accounts for roughly 40% of all DO tokens (e.g., Goldberg et al., 2004). On the other hand, a set of Latinate (i.e., fancy-sounding) verbs resist the DO in favor of the PO (Gropen et al., 1989; Ambridge et al., 2012; Goldberg, 2019). For instance, the verbs *transfer*, *explain,* and *donate* rarely occur in the DO, despite their highly compatible meanings; instead, each verb is biased toward the PO construction. Lexical biases can be quite particular and specific; for instance, the Latinate verb *guarantee*, bucks the tendency for fancy-sounding words to resist appearing in the DO: *Guarantee* strongly prefers the DO. Thus, a nuanced account of how such lexical factors are learned is required (e.g., Goldberg, 2011, 2019; Ambridge et al., 2012). Indeed, computational work has found that the differences in information structure between the DO and PO are useful in LLMs' learning of lexical biases (Misra and Kim, 2024).

As LLM representations are learned through exposure to natural language texts, there is an opportunity to investigate whether massive distributional learning can give rise to representations that reflect principles of the UCx approach. Recent work has assessed how accurately LLMs can classify or distinguish argument structure constructions (e.g. Huang, 2025; Bonial and Tayyar Madabushi, 2024), but less is known about *how* constructions are represented or their underlying geometry. Our work addresses this gap by shifting the focus from classification accuracy to an analysis of underlying representational geometry.

We hypothesize that representations of the DO and PO constructions should be more distinct to the extent that instances' typical lexical and functional properties are more prototypical instances of the respective constructions. We test this by asking whether collections of instances of the DO and PO that include typical functional features are more easily separable than collections of instances that are prototypical of neither, even though each sentence is unambiguous syntactically (either a PO or DO).

More specifically, we ask: Does the geometric distinction between the representations of the DO and PO increase, as measured by either energy distance or Jensen-Shannon Divergence, as the functional factors associated with DO and PO more closely align with their respective syntactic expressions?

Stimuli sentences come from the DAIS (Dative Alternation and Information Structure) dataset, which includes $5,000$ English pairs of DO and PO sentences (Hawkins et al., 2020). Across DO/PO pairs, several factors are systematically varied along the dimensions recognized to distinguish the two constructions. Here we use human preference strengths, also from DAIS, toward one or the other construction, to analyze the hidden states of Pythia-1.4B (Biderman et al., 2023).

Both energy distance (Rizzo and Székely, 2016) and Jensen-Shannon divergence (JSD) (Fuglede and Topsoe, 2004) are used to measure the separability of entire clouds of representations, at different layers of Pythia-1.4B. The preferred version (DO, PO, or either) of each sentence pair in the DAIS corpus was binned, according to the degree of preference toward the DO or PO, to be described in Methods and Results.

Results reveal a sophisticated geometric encoding of constructions in which lexical and functional factors improve the distinction between the DO and PO. In this way, LLM representations are consistent with key principles of the UCx approach, including gradiently distinguishable function-infused grammatical patterns, interpretable as clusters in geometric space.

2 Methods

Dataset and Model: As noted, the DAIS dataset includes 5000 pairs of sentences, one in the DO and one in the PO, while systematically varying the length and definiteness each postverbal argument across pairs. Two hundred main verbs also vary across pairs, including verbs standardly treated as both 'alternating' and 'non-alternating' (Levin, 1993). Importantly, DAIS also includes

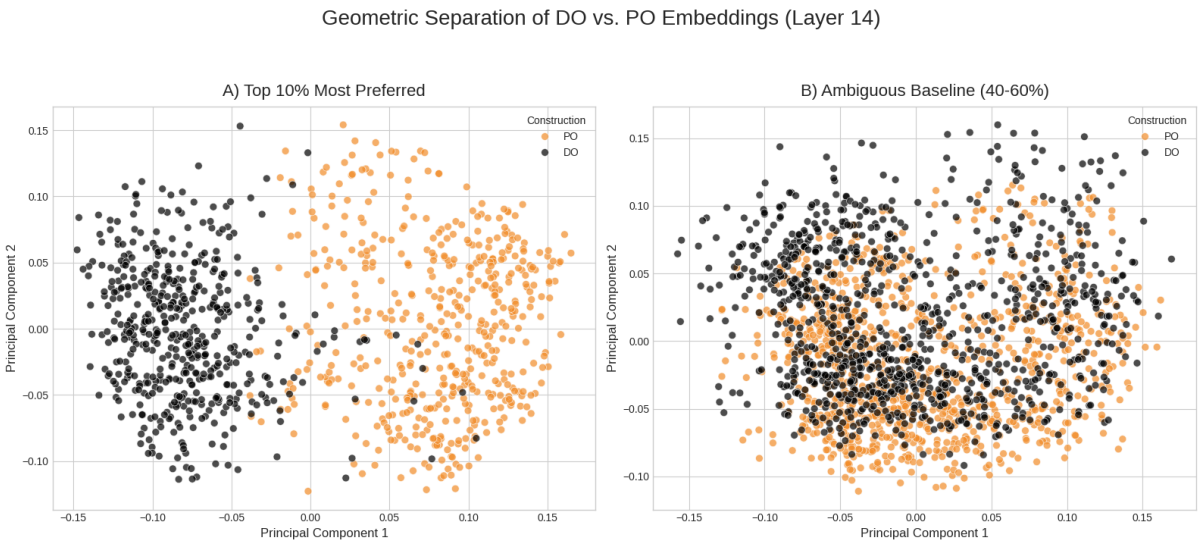


Figure 1: Projection into 2-dimensions of mean-pooled and normalized representations of the Double-Object (DO, in orange) and Prepositional Dative Object (PO, in black) constructions. Points represent sentences from the DAIS corpus, binned as follows: A) Instances of the DO and PO that are well-suited to the lexical and functional factors of their respective constructions as determined by human preferences. B) Instances of the DO and PO, as determined by syntax only, as their lexical and functional properties do not favor either construction.

human ratings of how strongly they prefer one construction over the other, for each combination of verb and arguments. Participants used a slider to indicate a preference for the DO (one end) or PO (other end) or neither (midpoint) (Hawkins et al., 2020).

We use these human preference ratings to partition sentences into five tiers based on the mean preference strength. We combined sentences from both ends of the scale to create 5 bins, ranging from: (1) the top 10% of sentences with the strongest preference for one or the other construction, to (5) those sentences judged to be in the middle of the scale (equally non-biased toward either construction). A sample collection of DO sentences that vary from strong DO-bias (1) to little bias toward either DO or PO (5) are provided below:

DO biased
ˆ (1) Maria asked him some questions.
| (2) Bob lobbed her a tennis ball.
| (3) Juan shuttled the team something.
| (4) Alice threw a woman a book.
| (5) Michael took the woman the blanket.
DO or PO

An equal number of PO sentences were included in each of the same bins, correspondingly ranging from strongly PO-biased (1) to equi-biased (5), according to the human preferences in DAIS.

From the publicly available pretrained Pythia-1.4B model, we extracted mean-pooled and normalized state representations for each sentence. We analyzed representations from all 24 layers, reducing them to 150 principal components, which captured 88.01% of the total variance (averaged across model layers). Common benchmarks suggest retaining components that explain 70%–90% of the total variance (Jolliffe, 2011), and 88.01% sits comfortably within this range, suggesting that a large majority of the structure in the data is retained, while a smaller portion that is more likely to reflect noise was discarded. We normalized the activations so that they all exist on a unit hypersphere S^{149}. Finally, we deployed the following analyses.

Preference strength was treated as an ordinal variable with five levels: 1 (10% most strongly biased) to 5 (10% most equi-biased). That is, level (1) includes sentences that strongly preferred the DO and sentences that strongly preferred the PO, while level (5) included sentences that were roughly equi-biased toward either DO or PO. Bins were used rather than a continuous factor for visualization purposes.

To measure the separability in representational space for each tier of bias strength, we em-

ployed two different measures, Energy Distance and Jensen-Shannon divergence. **Energy distance**, $\mathcal{E}(X, Y)$, is a statistical distance between the probability distributions of two random vectors, X and Y, in a metric space (Cramér, 1928). It is simply based on the expected Euclidean distances between their elements. Given two samples from our PCA-reduced representations, $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, where each $x_i, y_j \in \mathbb{R}^{150}$, the squared energy distance is estimated as:

$$\mathcal{E}^2(X,Y) = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|x_i - y_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|x_i - x_j\| \quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|y_i - y_j\|$$

where $\|\cdot\|$ is the Euclidean norm. Energy distance is zero if and only if the distributions are identical; it is sensitive to differences in both the location and the shape of distributions, making it a robust measure of overall geometric separation in the model's representation space. We calculated the energy distance between the distributions of the constructions on S^{149} layer by layer.

A more sensitive measure of distributions is **Jensen-Shannon Divergence (JSD)**, which measures the relationship between distributions in high-dimensional space (Menéndez et al., 1997). To use JSD, we first estimated the probability distributions of both constructions $P(v_{DO})$ and $Q(v_{PO})$, in the 150-dimensional PCA space. Following (Conklin, 2025), we next generated a set of $k = 1000$ *anchor vectors*, $A = a_1, a_2, \ldots, a_k$, by sampling from a uniform distribution on S^{149}. The vector corresponding to each individual sentence v was then assigned to the anchor vector nearest to it, based on cosine similarity. This effectively partitions the hypersphere into k Voronoi cells. This technique, based on vector quantization, yields two discrete probability distributions, $\hat{P}$ and $\hat{Q}$, which are k-dimensional vectors where the i-th element represents the proportion of vectors from each set assigned to anchor a_i. Jensen-Shannon divergence is then computed as:

$$JSD(\hat{P}\|\hat{Q}) = \frac{1}{2}D_{KL}(\hat{P}\|M) + \frac{1}{2}D_{KL}(\hat{Q}\|M)$$

where $M = \frac{1}{2}(\hat{P} + \hat{Q})$ and D_{KL} is the Kullback-Leibler divergence. This method avoids information loss from projecting onto any single axis to offer a more holistic comparison of distributions (Conklin, 2025). Since the sampled anchor vectors are probabilistic, we averaged across 20 random seeds to get stable JSD scores.

3 Results

Our analysis reveals that graded bias strength for one construction over a paraphrase systematically shapes the geometry of construction representations across the model architectures. In particular, the model assigns representations that are more distinct when the constructions are more clearly differentiated, when instances are more strongly biased toward the construction used. This is the case for both energy distance and JSD, as each shows a clear and consistent stratification by the tiers of preference strength (Figure 2 and 3). At nearly every layer, the Top 10% strongest preference tier exhibits the greatest geometric distance, followed in order by the other tiers, down to the ambiguous baseline. As is clear in Figure 1, with sentence vectors projected onto two dimensions for visualization purposes, DO and PO sentences that conform better to the DO or the PO, respectively (left panel), are more distinctive than sentences that could nearly as easily be paraphrased by the other construction (right panel).

Because energy distance and JSD are based on very different analyses, we cannot expect their qualitative patterning to align. In fact, energy distance follows a convex trajectory, slightly dipping in the mid-layers before rising sharply (Figure 2) In contrast, JSD shows divergence increasing sharply and remaining high (Figure 3). Yet the overall pattern showing more distinctiveness between DO vs. PO sentences when sentences that are more biased toward either varient compared with sentences that are relaitvely un-biased is evident according to both energy distance (Figure 3) and JSD (Figure 2). We take this to indicate that the finding is robust and not an artifact of a single metric.

Scaling Analysis across Pythia Model Suite. To test whether our findings are specific to the 1.4B parameter model or reflect a more general property of transformer architectures, we replicated our full analysis pipeline — including the energy distance, high-dimensional JSD with correlation of mean cosine similarity tests — across the models in the Pythia suite (from 70M to 6.9B parame-

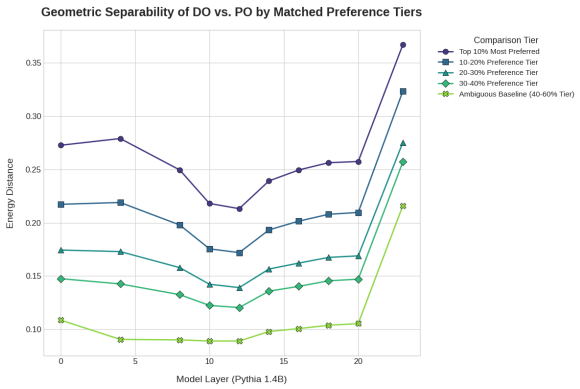


Figure 2: Layerwise Energy Distance between DO and PO representations, stratified by tiers binned by degree of bias. The plot shows a consistent ordering by bias, with more prototypical instances of the two constructions being more separable.

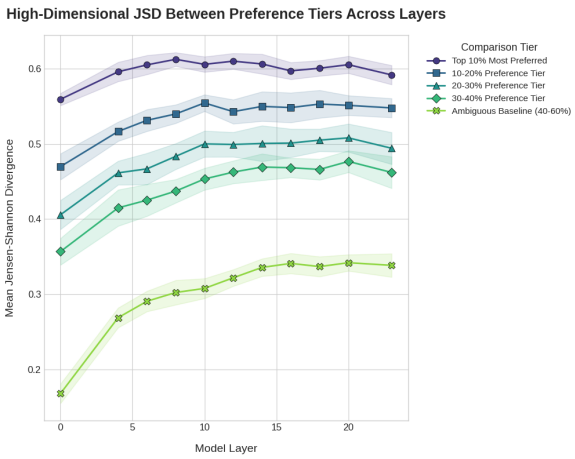


Figure 3: A high-dimensional measure of Distributional Separability (JSD) Across layers (with $k = 1000$ anchor points, averaged over 20 random seeds). Stratification is evident: tiers that include more biased sentences of either DO or PO are more separable into distinct constructions. This plot reinforces the finding from the energy distance analysis, though with different layer-wise dynamics.

ters). Results confirm that our central findings are robust across model scales. We consistently observe the geometric stratification by degree of bias so that preference strength remains a significant predictor of representational distance. A detailed report of these scaling law analyses, with code to generate plots, is in the supplementary materials.

4 Discussion

Our results provide compelling computational evidence for a core principle of the usage-based constructionist approach: that grammatical representations are graded and sensitive to semantic-pragmatic fit. The clear stratification in our geometric analyses demonstrates that LLMs develop representations whose geometric properties are highly consistent with the probabilistic, usage-based categories posited by the UCx approach. This geometric entanglement of form and function resonates with the core tenets of the UCx approach. The energy distance reflects the distinctiveness of the two constructions spatially in regions of the model's representation space. This extends previous work that has focused on the model's ability to classify constructions categorically (Huang, 2025; Bonial and Tayyar Madabushi, 2024) by showing more fine-grained, graded geometric structure, dependent on lexical and functional factors.

Future work is needed to better understand the distinct qualitative patterns across layers when energy distance and JSD are compared. We note that the JSD measure, which is more nuanced but perhaps less intuitive, appears to distinguish the constructions particularly well: JSD is bounded between 0 and log(2) (≈ 0.693) (Lin, 1991), and the distinction between the constructions in the most biased tier approaches this limit at 0.6. Yet because the two metrics are based on quite different calculations, so we do not attempt to compare them directly.

5 Conclusion and future directions

The current work demonstrates that the geometry of an LLM's internal representations directly reflects the graded function-infused bias toward one or another linguistic construction, where biases are recognized to be conditioned on lexical and functional factors. We have shown that the model's representations of constructions are systematically organized by their distinctiveness. This work bridges the gap between the theoretical principles of the usage-based constructionist approach and the empirical realities of modern NLP, suggesting that LLMs learn a rich, dynamic, and meaning-infused model of grammar. Our findings open a promising new direction for future work; we are currently using these geometric insights to guide an investigation aimed at isolating the specific computational circuit(s) within the model that are responsible for encoding verb bias in the dative alternation, using tools like causal mediation analysis from Mechanistic Interpretability.

Acknowledgements

We are grateful to the three anonymous reviewers for their insightful comments and constructive feedback, which significantly helped clarify and strengthen the arguments presented in this paper.

References

Ben Ambridge, Julian M Pine, Caroline F Rowland, and Franklin Chang. 2012. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language*, 88(1):45–81.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2217–2256. PMLR.

Claire Bonial and Harish Tayyar Madabushi. 2024. Constructing understanding: on the constructional information encoded in large language models. *Language Resources and Evaluation*, pages 1–40.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In Gosse Bouma, Irene Kraemer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences, Amsterdam.

Joan Bresnan and Tatiana Nikitina. 2009. The gradience of the dative alternation. *Reality exploration and discovery: Pattern interaction in language and life*, pages 161–184.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. *arXiv preprint arXiv:2305.18598*.

Henry Conklin. 2025. Information structure in mappings: An approach to learning, representation, and generalisation. *arXiv preprint arXiv:2505.23960*.

Harald Cramér. 1928. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Elaine Francis. 2022. *Gradient acceptability and linguistic theory*, volume 11. Oxford University Press.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

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

Adele E Goldberg. 2002. Surface generalizations: An alternative to alternations.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, New York.

Adele E Goldberg. 2011. Corpus evidence of the viability of statistical preemption.

Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Adele E Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

Adele E Goldberg, Devin M Casenhiser, and Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive linguistics*, 15(3):289–316.

Georgia M Green. 1974. Semantics and syntactic regularity. bloomington. *(No Title)*.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in english. *Language*, pages 203–257.

Robert D. Hawkins, Ngan Nguyen, Adele E. Goldberg, Michael C. Frank, and Noah D. Goodman. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4718, Online. Association for Computational Linguistics.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Haerim Huang. 2025. Assessing the performance of pretrained models for accurate and consistent classification of argument structure constructions. In *Applied Artificial Intelligence*.

I. Jolliffe. 2011. Principal component analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer, Berlin, Heidelberg.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C. Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models: A case study on cross-dative generalization. *arXiv preprint arXiv:2408.05086*.

Richard Thomas Oehrle. 1976. *The grammatical status of the English dative alternation*. Ph.D. thesis, Mass. Cambridge.

J. Ranganathan, R. Jha, K. Misra, and K. Mahowald. 2025. semantic-features: A user-friendly tool for studying contextual word embeddings in interpretable semantic spaces. ArXiv preprint arXiv:2506.XXXXX.

Maria L Rizzo and Gábor J Székely. 2016. Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38.

Thomas Wasow and Jennifer Arnold. 2003. Post-verbal constituent ordering in english. *Determinants of grammatical variation in English*, pages 119–154.

Annotating English Verb-Argument Structure via Usage-Based Analogy

Allen Minchun Hsiao
University of Colorado Boulder
Min-Chun.Hsiao@colorado.edu

Laura A. Michaelis
University of Colorado Boulder
Laura.Michaelis@colorado.edu

Abstract

This paper introduces a usage-based framework that models argument structure annotation as nearest-neighbor classification over verb–argument structure (VAS) embeddings. Instead of parsing sentences separately, the model aligns new tokens with previously observed constructions in an embedding space derived from semi-automatic corpus annotations. Pilot studies show that cosine similarity captures both form and meaning, that nearest-neighbor classification generalizes to dative alternation verbs, and that accuracy in locative alternation depends on the corpus source of exemplars. These results suggest that analogical classification is shaped by both structural similarity and corpus alignment, highlighting key considerations for scalable, construction-based annotation of new sentence inputs.

1 Introduction

Verbs provide a crucial interface between syntax and semantics, typically determining both the nature of the event or action described by a clause and the number and type of participants the clause contains—a configuration known as **argument structure**. For example, the verb *give* denotes an act of possession transfer and therefore requires three arguments: an `AGENT` (*Paul*), a `RECIPIENT` (*me*), and a `THEME` (*a book*). These may be realized syntactically as the Double Object construction, as in *Paul gave me a book*. At the same time, proponents of construction-based syntax have observed that verbs may be mismatched to their syntactic contexts in ways that alter the meaning and valence of the verb. For example, while creation verbs like *paint* and *draw* do not intrinsically express acts of transfer, they can be used to implicate actual or intended transfer in sentences like *I drew her a picture*. Such examples suggest that argument-structure patterns themselves can convey event structures traditionally attributed to verbs alone and, in turn, may influence the verb's meaning and selectional properties (Goldberg, 1995; Michaelis, 2004).

To obtain argument structures from natural language, most NLP systems rely on automatic Semantic Role Labeling (SRL) and constituency parsing. SRL identifies argument spans and assigns semantic roles (Màrquez et al., 2008; Gildea and Jurafsky, 2002), while constituency parsing extracts hierarchical phrase structures (Marcus et al., 1993). However, these two components are often modeled separately, leading to cascading errors: syntactic misparses can degrade SRL accuracy. Recent approaches integrate syntactic information into neural SRL models (Strubell et al., 2018; Zhou et al., 2020; Fei et al., 2021), and BERT-based architectures frame SRL as span classification without explicit syntactic features (Shi and Lin, 2019). Meanwhile, high-accuracy constituency parsers like the Berkeley Neural Parser (Kitaev and Klein, 2018) continue to be widely used in such pipelines.

Yet despite their strong performance, these systems are optimized for SRL as a classification task and maintain a verb-centered view of argument structure. Even models that integrate syntax typically do so only to improve SRL performance. Moreover, most SRL datasets, such as PropBank and VerbNet, rely on verb-specific argument frames, which limit generalization across constructions. Arguments introduced by constructions—rather than verbs—are often overlooked. For instance, in "*kick the ball into the room*," the directional PP *into the room* is typically treated as an adjunct, despite fulfilling a core semantic role (`Goal`) in a Caused-Motion construction. As a result, these models fall short of capturing the full range of construction-based argument structures observed in natural language.

This paper introduces a usage-based alternative that models argument structure through analogical matching against previously observed VAS patterns. Instead of parsing a sentence and mapping its elements via fixed templates, our model compares the sentence to a library of VAS exemplars and selects the best match in embedding space. This nearest-neighbor approach treats argument structure as a

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 158–164

product of linguistic experience, and contextual inference, rather than static verb valency.

The remainder of this paper is organized as follows: Section 2 reviews Construction Grammar-based approaches to argument structure annotation; Section 3 introduces our model; Section 4 details the methodology, including data curation and annotation; Section 5 reports pilot studies; and Section 6 concludes with discussion and future directions.

2 Related Work

Resources grounded in Construction Grammar (CxG) aim to annotate argument structures that arise not only from lexical valency but also from constructional licensing. Kyle and Sung (2023) introduce the first argument structure construction (ASC) treebank, manually annotating verb-argument structures following CxG principles. While valuable, the treebank covers only a small set of constructions, limiting its generalizability.

Perek and Patten (2019) explore the empirical identification of constructions using syntactic n-grams extracted from the British National Corpus. They cluster these "treelets" by distributional similarity and manually select a linguistically meaningful subset. This work lays a data-driven foundation for construction identification in English but requires extensive manual intervention and remains a work in progress.

Computational frameworks like Fluid Construction Grammar (FCG) (Beuls and Van Eecke, 2023) offer a cognitively motivated architecture for representing argument structure via learned form-meaning pairings. FCG supports both parsing and production, modeling the dynamic invocation of constructions during language use. However, despite its expressive power, FCG relies on hand-engineered constructions and operates primarily in simulation environments or controlled domains. It lacks scalable interfaces with large corpora or pretrained models. As a result, while FCG demonstrates the theoretical utility of construction-based approaches, it is not yet suitable for automatically annotating verb-argument structures in real-world corpora. Because of this limitation, we continue to seek scalable, data-driven alternatives.

3 Our Model

This project introduces a usage-based, analogy-driven framework for annotating VAS in natural language. Drawing on exemplar-based approaches to grammar (Bybee, 2013), the model treats argument structure annotation as a **nearest-neighbor retrieval task** grounded in **analogical matching** (Gentner, 1983, 2010). Given a sentence containing a target verb, it selects the most likely structure by comparing the verb's contextual embedding to a set of previously observed, type-level verb-argument structure embeddings. In essence, it asks: "Which known pattern does this usage most closely resemble?"

Rather than assuming each verb is tied to a fixed valency frame, the model is built on the idea that argument structures generalize across verbs. For example, the structure associated with *give*, as in *She gave him a book*, may serve as an exemplar for annotating *bequeath*, as in *She bequeathed him a book*. Such generalizations are achieved through analogical matching: the model ranks known structures by their cosine similarity to the target usage, offering plausible annotations even for novel or infrequent verbs.

Each verb-argument structure in the model is represented as an embedding derived from actual corpus attestations. These type-level embeddings are stored and compared against token-level embeddings extracted from new input sentences. The top-ranked structures—those most similar in both form and meaning—are returned as candidate annotations. This ranked list supports both automatic labeling and human-in-the-loop annotation, functioning as an assistant that provides interpretable and transparent suggestions.

4 Methodology

This section outlines the steps used to develop our model, from data selection and annotation to the construction of a semantic space of verb-argument structures (VAS). We divide the methodology into three components: (1) data curation, (2) verb-argument structure annotation, and (3) construction of the VAS space.

4.1 Data Curation

Corpus Selection. We use a subset of the BabyLM Project Gutenberg corpus (Warstadt et al., 2023), which contains written English texts from books in the public domain. Our goal is not exhaustive annotation but the development of a representative VAS space from high-quality language data. We sample 51,411 sentences for analysis.

Data Filtering. To focus on clause-level predicates, we filter the data using spaCy (Honnibal et al., 2020). We retain only sentences where the main verb is the syntactic `ROOT` and has a nominal subject (`NSUBJ`). This reduces the dataset to 30,139 sentences. We further exclude malformed or fragmentary sentences, yielding a final dataset of 23,396 well-formed sentences.

4.2 Verb-Argument Structure Annotation

Initial semantic-syntactic auto-annotation. We begin by automatically annotating each verb-argument structure with syntactic phrase types and semantic roles. Semantic roles are assigned using SemParse (Gung, 2020), which maps predicates to VerbNet classes and extracts PropBank-style arguments; these are then converted to VerbNet roles using the mappings in Kipper-Schuler et al. (2008). Syntactic categories are obtained from the Berkeley Neural Parser (Kitaev and Klein, 2018), from which we extract the highest syntactic projection of each argument. These initial annotations are used as a base for further revision.

Construction-Based Revision. Following principles in Construction Grammar (Goldberg, 2006; Michaelis, 2012), we revise the initial annotations to reflect arguments introduced not only by the verb's lexical valency but also by larger constructions. For example, in the sentence *She kicked the ball into the room*, the directional phrase *into the room* is labeled as a `Goal` argument—not as an adjunct—because it is licensed by the Caused-Motion construction rather than by the verb *kick* alone. These construction-based revisions ensure that the final annotations more accurately capture the full range of argument structure patterns observed in natural usage.

4.3 Constructing the VAS Space

Embedding Extraction. Each verb-argument structure instance is represented as a contextualized embedding extracted from BERT (Devlin et al., 2019), using layer 7 (Chronis and Erk, 2020). Embeddings are grouped by VAS type and averaged to yield a type-level embedding.

Linguistic Experience Space. The resulting VAS space serves as a structured repository of linguistic experience. Each type-level embedding encodes the distributional and constructional properties of a verb-argument structure as observed in corpus data. Given a new sentence, the model retrieves the most similar structure in the space using cosine similarity. This usage-based approach reflects how speakers interpret novel utterances by analogizing to familiar patterns encountered in prior language use.

5 Pilot Studies

This section reports three pilot studies. The first examines what cosine similarity between verb embeddings captures. The second and third test how well the model can classify unseen verb tokens using a small set of precomputed structure embeddings.

5.1 Pilot 1: What Does Cosine Similarity Capture?

The first pilot study examines what cosine similarity between verb token embeddings captures, since this similarity metric underlies our method for selecting candidate structures. Understanding what it reflects—**surface form** (i.e., syntactic realization), **relational meaning** (i.e., argument structure roles), or **both**—is essential for evaluating the validity of our analogical classification approach.

We test this using the verb *bequeath*, which alternates between the **Prepositional Dative** and **Double Object** constructions. For each form, we compare its embedding to three minimally altered sentences, varying only the verb while holding the surrounding context constant. This isolates the verb's syntactic and semantic contribution as the source of variation in cosine similarity.

Double Object variant:

The widow bequeathed the church her property.
<NP1$_{\text{Agent}}$, NP2$_{\text{Recipient}}$, NP3$_{\text{Theme}}$>

- *The widow gave the church her property.* (form and meaning) → cosine similarity: 0.7541
- *The widow gave her property to the church.* (meaning only) → cosine similarity: 0.7243
- *The widow considered the church her property.* (form only) → cosine similarity: 0.5481

Prepositional Dative variant:

The widow bequeathed her car to the church.
<NP1$_{\text{Agent}}$, NP2$_{\text{Theme}}$, PP$_{\text{Recipient}}$>

- *The widow gave her car to the church.* (form and meaning) → cosine similarity: 0.7753
- *The widow gave the church her car.* (meaning only) → cosine similarity: 0.7119

- *The widow drove her car to the church.* (form only) → cosine similarity: 0.5962

In both constructions, the sentence sharing both form and meaning with the target had the highest similarity, while form-only matches scored lowest. Meaning-only matches consistently ranked in between. This pattern indicates that **cosine similarity in our embedding space captures both syntactic and semantic similarity, with a stronger bias toward meaning**. These results support the use of cosine similarity for analogical classification and align with the usage-based view that novel utterances are understood through semantic alignment with familiar constructions.

5.2 Pilot 2: Nearest-Neighbor Classification Accuracy on Dative Alternation Verbs

To test our model's classification accuracy, we evaluated whether unseen verb tokens could be correctly assigned argument structure labels by comparing their embeddings to five precomputed VAS embeddings of *give*. We tested both **within-verb generalization**—predicting new *give* tokens drawn from COCA—and **cross-verb generalization**—predicting an **unseen** verb *bequeath* from COCA.

5.2.1 Experimental Setup

We manually annotated five distinct VAS types for *give* from the Gutenberg corpus (see Appendix A.1). Each structure combined specific semantic roles and phrase types and served as a prediction template.

The test set included 40 sentences from COCA: 20 with *give* and 20 with *bequeath*, each evenly split between the Double Object and Prepositional Dative constructions. Each verb token was embedded using BERT (layer 7) and matched to the most similar *give* structure embedding based on cosine similarity.

5.2.2 Results and Discussion

Table 1 shows that the model achieved high accuracy across both within-verb and cross-verb conditions. For *give*, which appeared in the training corpus (Gutenberg), the model correctly predicted all 10 Double Object tokens and 9 out of 10 Prepositional Datives from the test set (COCA). For *bequeath*, which was unseen during training, the model correctly classified all 20 tokens.

These results suggest that our model can generalize both to new uses of familiar verbs and to entirely new verbs that share similar constructions. The success of cross-verb classification, especially for a rare verb like *bequeath*, indicates that the precomputed structure embeddings of *give* encode transferable, construction-level information. This supports our central hypothesis: **verb-argument structure annotation can be modeled as a nearest-neighbor classification task in a semantically structured space**.

Verb	Argument Structure	Accuracy
give	Double Object	100% (10/10)
	Prepositional Dative	90% (9/10)
bequeath	Double Object	100% (10/10)
	Prepositional Dative	100% (10/10)

Table 1: Nearest-neighbor classification accuracy for *give* and *bequeath* using *give* structure embeddings.

5.3 Pilot 3: Nearest-Neighbor Classification Accuracy on Locative Alternation Verbs

We next tested our model on the verb *spray*, which alternates between the Caused-Motion (CM) construction (e.g., *He sprayed the paint onto the wall*) and the Theme-Applicative (TA) construction (e.g., *He sprayed the wall with paint*). This alternation provides an ideal case study because it involves two competing argument structure frames that are both frequent and semantically transparent, yet distinct in terms of syntactic realization.

Experimental setup. We assembled 100 tokens of *spray*, balanced between 50 CM and 50 TA tokens. Of these, 20 CM and 30 TA tokens were drawn directly from COCA, and each was paired with an altered counterpart in the alternate construction (e.g., a TA token such as *Pinocchio sprays Puss with water* was paired with its CM variant *Pinocchio sprays water to Puss*). This procedure yielded a balanced dataset where every naturally attested token was matched with a constructed counterpart, ensuring equal representation of both constructions.

Two VAS type embeddings served as classifiers. For the CM frame, we used the *put* pattern (*He put the money into the pocket*), averaged from 66 CM *put* tokens in Gutenberg. For the TA frame, we used the *cover* pattern (*He covered his beard with his hands*), averaged from 50 TA tokens in COCA.[1] Both *put* and *cover* serve as prototypical

[1]We did not use the TA *cover* pattern from Gutenberg due

exemplars of their respective constructions, making them suitable analogical anchors for classification.

Results with Gutenberg *put*. When paired against the COCA *cover* embedding, the Gutenberg *put* embedding produced an accuracy of 0.690 and a macro F1 of 0.662 (Table 4). Predictions were heavily skewed toward the TA category: of 100 spray tokens, 79 were classified as TA, yielding an F1 of 0.760 for TA but only 0.563 for CM. The full confusion matrix is shown in Table 2.

Gold ARGST	Predicted: TA	Predicted: CM
TA	**49**	1
CM	**30**	20

Table 2: Confusion matrix for *spray* prediction using Gutenberg *put* (CM) vs. COCA *cover* (TA).

This imbalance suggested that the skew might not be due to structural similarity alone, but instead to corpus mismatch: both *spray* and the TA source (*cover*) came from COCA, while the CM source (*put*) came from Gutenberg. To test this speculation, we repeated the experiment using a CM *put* embedding drawn from COCA rather than Gutenberg.

Results with COCA *put*. Substituting 50 CM *put* tokens from COCA yielded stronger performance: accuracy rose to 0.860 and macro F1 also reached 0.860 (Table 4). Predictions were more balanced, with F1 scores of 0.865 for TA and 0.854 for CM. The corresponding confusion matrix is shown in Table 3.

Gold ARGST	Predicted: TA	Predicted: CM
TA	**45**	5
CM	9	**41**

Table 3: Confusion matrix for *spray* prediction using COCA *put* (CM) vs. COCA *cover* (TA).

Discussion. Taken together, the results summarized in Table 4 suggest that analogical classification may be affected by the corpus where the source patterns are drawn. When sources and targets were drawn from different corpora (Gutenberg vs. COCA), predictions skewed heavily toward the COCA source. When both sources were drawn from COCA, predictions became more balanced and overall accuracy improved. Although these findings are preliminary, they indicate that corpus alignment could interact with structural similarity in shaping analogical predictions. Future work will test this more systematically across additional verbs, constructions, and corpora.

	CM-*put* (Gut.) + TA-*cover* (COCA)	CM-*put* (COCA) + TA-*cover* (COCA)
Accuracy	0.690	**0.860**
Macro F1	0.662	**0.860**
F1 (TA)	0.760	**0.865**
F1 (CM)	0.563	**0.854**

Table 4: Summary of nearest-neighbor prediction performance for *spray*.

6 Conclusion

We presented a usage-based model that operationalizes argument structure annotation as a nearest-neighbor classification task over verb–argument structure (VAS) embeddings. By aligning new sentences with previously encountered constructions in a multidimensional embedding space, the model reflects how speakers interpret novel expressions—not by parsing syntax and semantics separately, but by recognizing patterns grounded in prior linguistic experience.

Our pilot studies illustrate both the promise and the challenges of this approach. Pilot 1 showed that cosine similarity captures differences in both form and meaning, with a stronger bias toward meaning. Pilot 2 demonstrated that nearest-neighbor classification can model argument structure in dative alternation verbs, and together with Pilot 3, showed that a single VAS type embedding can support accurate prediction across verbs. Pilot 3 further scaled up to locative alternation verbs and revealed that accuracy also depends on corpus source: predictions were more accurate when sources and targets came from the same corpus. These findings suggest that analogical classification is shaped not only by structural similarity but also by corpus alignment, pointing to key considerations for future large-scale work.

The framework is designed for continuous refinement: new structure types and attestations can be added over time, allowing it to evolve alongside linguistic theory and empirical data. This scalability supports interpretable annotation and underscores the value of high-quality, construction-based analysis—even in the era of large embedding models.

to its limited size ($n = 13$).

References

Katrien Beuls and Paul Van Eecke. 2023. Fluid construction grammar: State of the art and future outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 41–50, Washington, D.C. Association for Computational Linguistics.

Joan L. Bybee. 2013. Usage-based theory and exemplar representations of constructions. In *The Oxford Handbook of Construction Grammar*, volume 1. Oxford University Press.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. 2021. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, Online. Association for Computational Linguistics.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Dedre Gentner. 2010. Bootstrapping the mind: analogical processes and symbol systems. *Cognitive Science*, 34(5):752–775.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford, UK.

James Gung. 2020. SemParse, VerbNet Parser.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Karin Kipper-Schuler, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Kristopher Kyle and Hakyung Sung. 2023. An argument structure construction treebank. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 51–62, Washington, D.C. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Special issue introduction: Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Laura A. Michaelis. 2004. Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive Linguistics*, 15(1):1–67.

Laura A. Michaelis. 2012. Making the case for construction grammar. In Hans C. Boas and Ivan A. Sag, editors, *Sign-Based Construction Grammar*, pages 31–69. CSLI Publications, Stanford, CA.

Florent Perek and Amanda Patten. 2019. Towards an inventory of english argument structure constructions: Empirical assessment of the n-gram approach. *Constructions and Frames*, 11(2):250–282.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*. Work in progress.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus.

Junru Zhou, Zuchao Li, and Hai Zhao. 2020. Parsing all: Syntax and semantics, dependencies and spans. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4438–4449, Online. Association for Computational Linguistics.

Appendix

A.1 Verb-Argument Structure Types for *give*

Below are the five verb-argument structure types of *give* used in the classification model, each accompanied by an illustrative sentence from the Gutenberg corpus:

1. <NP1$_{\text{Agent}}$, NP2$_{\text{Recipient}}$, NP3$_{\text{Theme}}$>
   *They **gave** him an opportunity of speaking more, and therefore he thought himself better than the rest.*

2. <NP1$_{\text{Agent}}$, NP2$_{\text{Recipient}}$, S[QUE+]$_{\text{Theme}}$>
   *Simonetta **gave** her mother what was indispensable for household expenses and managed the rest herself.*

3. <NP1$_{\text{Agent}}$, NP2$_{\text{Theme}}$, PP$_{\text{Recipient}}$>
   *He should have **given** the deer to the woman.*

4. <NP1$_{\text{Agent}}$, NP2$_{\text{Theme}}$, PP$_{\text{Beneficiary}}$>
   *With the same humanity which they had shown in the case of Jogues, they **gave** a generous ransom for him, supplied him with clothing, kept him until his strength was in some degree recruited, and then placed him on board a vessel bound for Rochelle.*

5. <NP1$_{\text{Agent}}$, NP2$_{\text{Theme}}$>
   *[...] the magician **gives** the order for preparations.*

Can Constructions "SCAN" Compositionality ?

Ganesh Katrapati and **Manish Shrivastava**
International Institute of Information Technology Hyderabad
ganesh.katrapati@research.iiit.ac.in m.shrivastava@iiit.ac.in

Abstract

Sequence to Sequence models struggle at compositionality and systematic generalisation even while they excel at many other tasks. We attribute this limitation to their failure to internalise *constructions*—conventionalised form–meaning pairings that license productive recombination. Building on these insights, we introduce an unsupervised procedure for mining *pseudo-constructions*: variable-slot templates automatically extracted from training data. When applied to the SCAN dataset, our method yields large gains out-of-distribution splits: accuracy rises to **47.8%** on ADD JUMP and to **20.3%** on AROUND RIGHT without any architectural changes or additional supervision. The model also attains competitive performance with $\leq 40\%$ of the original training data, demonstrating strong data efficiency. Our findings highlight the promise of construction-aware preprocessing as an alternative to heavy architectural or training-regime interventions.

1 Introduction

Compositionality is the principle that the meaning of a complex expression is determined by the meanings of its parts and the rules used to combine them (Fodor and Pylyshyn, 1988; Marcus, 2003; Partee et al., 1990). It enables systematic generalisation: the ability to understand and produce novel combinations of familiar elements, a hallmark of human language competence.

Despite the impressive empirical performance of sequence to sequence models such as RNNs, LSTMs, and Transformers, studies have consistently found that they struggle with tasks requiring compositional generalisation (Lake and Baroni, 2018; Hupkes et al., 2020; Keysers et al., 2020). When faced with inputs that combine known primitives in unseen ways, these models frequently fail to extrapolate correctly.

Cognitive and Construction Grammar treat *constructions* as form–meaning pairs composed of conventionalised components that combine with lexical items (Goldberg, 1995; Langacker, 1987; Croft, 2001). For successful communication, speakers must have access to these conventionalised constructions shared within their linguistic community. The degree of conventionalisation varies across construction types: for example, idiomatic expressions like "kick the bucket" are fully fossilised and resist internal modification, whereas partially filled constructions such as "the Xer the Yer" contain open slots that can be flexibly filled to produce complete surface forms (Fillmore et al., 1988; Goldberg, 2006).

Inspired by this notion, we propose that modelling constructions is essential to solving the problem of compositionality. We choose the SCAN dataset - a canonical testbed for evaluating compositionality in neural models - to demonstrate our approach. We introduce a simple yet effective method of mining *pseudo-constructions* and show that models trained on segmented data achieve significant improvements over standard baselines on SCAN's ADD JUMP and AROUND RIGHT splits.

Furthermore, we demonstrate strong data efficiency: by leveraging the compositional structure, our method requires substantially less data to achieve competitive performance, especially on simpler splits. Our results suggest that carefully exposing compositional patterns during training can yield robust improvements without resorting to complex interventions.

2 Related Work

There have been a number of benchmarks and tasks to evaluate whether modern NLP methods including deep neural networks such as RNNs (Elman, 1990), LSTMs (Hochreiter and Schmidhuber,

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 165–171

1997) and Transformers (Vaswani et al., 2017) exhibit compositional behaviour. *SCAN* (Lake and Baroni, 2018), *COGS* (Kim and Linzen, 2020) , *CFQ* (Keysers et al., 2020), *PCFG* (Hupkes et al., 2020) and similar benchmarks focus on sequence prediction tasks where input sequence must be processed in a compositional manner to yield the correct sequence on the target side.

They showed that the models do *not* generalise systematically: when confronted with new combinations of words or phrases that were absent from the training data, their performance breaks down. Subsequent studies on a variety of datasets (Li and colleagues, 2021; Sinha et al., 2019; Liška et al., 2018), have reported similar findings. Informed by these limitations, recent work has led to multiple methods to improve compositional generalisation abilities of neural network models.

Multiple studies have focused on disentangling syntax and semantics - (Russin et al., 2019) introduced a dedicated syntactic channel boosts SCAN accuracy dramatically , separating primitive–function pathways pushes performance to near-perfect levels (Li et al., 2019; Jiang and Bansal, 2021). Rather than separating syntax and semantics, some studies have focused on syntactic guidance. (Hupkes et al., 2019; Baan et al., 2019; Kim et al., 2021; Zanzotto et al., 2020).

Data-centric approaches improve compositionality by augmenting the training corpus with systematically recombined examples: GECA (Andreas, 2020), automatically mined lexical symmetries (LEXSYM; Akyürek and Andreas, 2022), and grammar-based generators such as CSL (Qiu et al., 2022) all substantially cut error rates on SCAN, COGS, and CLEVR. Herzig et al. (2021) insert a reversible or lossy intermediate representation between the input and the target program, doubling accuracy on CFQ MCD splits and adding 15–20 points on text-to-SQL.

Treating compositionality as a transferable skill, Meta learning approaches (Zhu et al., 2021; Lake, 2019; Lake and Baroni, 2023) push transformers beyond 70 % accuracy on the hardest SCAN and COGS splits.

Apart from this, several studies have proposed significant modifications to the neural network architecture (Csordás et al., 2022; Huang et al., 2024) and neural–symbolic designs such as NMN, MAC, NLM, LANE, program-synthesis grammars, and the Neural-Symbolic Recursive Machine (Andreas et al., 2017; Hudson and Manning, 2018; Dong et al., 2019; Liu et al., 2022; Nye et al., 2020; Li et al., 2022) which achieve (near-)perfect compositional generalisation on datasets like SCAN, COGS and CFQ.

While many of these approaches achieve near-perfect accuracy in datasets like SCAN and COGS, they either require data augmentation, which likely translates into training bigger models for a longer time, or they propose drastic architectural changes which have not been proven to scale beyond these benchmarks. Our method does not employ data augmentation or complex architectural changes. Our aim is show that taking insights from Cognitive Grammar and the notion of *Constructions* leads to building models more capable of compositional generalisation.

Recent work on integrating Construction Grammar (CxG) with neural models has been encouraging: fine-tuning BERT on construction-annotated corpora sharpens its encoding of construction identity and slot fillers (Tayyar Madabushi et al., 2020), a Mandarin CxLM leverages more than ten-thousand schemata to boost cloze accuracy (Tseng et al., 2022). Yet no study has directly shown that construction-aware training itself improves systematic compositional generalisation on classic out-of-distribution tests and bridging that gap remains a challenge.

3 Data

Introduced by (Lake and Baroni, 2018), *SCAN* contains pairs of simple navigation commands with action sequences; primitives like “jump” map to “`I_JUMP`”, while modifiers such as “left”, “right”, “opposite”, and “around” compose these primitives into longer actions.

The original paper showed that models excel on a random split yet falter on novel combinations. In the ADD JUMP split, models see the primitive “jump” during training but must execute composed forms (e.g., “jump twice”) at test time. Loula et al. (2018) extended this with the AROUND RIGHT split: training includes “walk left”, “walk right”, “jump around left”, and so on, while testing requires generalising to “jump around right”, forcing the model to learn that “around” modifies directions and that “left” and “right” are symmetric.

We focus on improving the accuracy for both these splits.

4 Approach

Definition (Pseudo-construction). A *pseudo-construction* is a partially specified template induced from training data, containing fixed words alongside one or more *slots* represented by placeholders (e.g., _ or W_n). Unlike fully conventionalised constructions, pseudo-constructions are derived automatically and capture recurring structural patterns that can generalise to novel inputs when the slots are filled with appropriate lexical items.

4.1 Mining Pseudo-constructions

A SCAN train or test set consists of both a source file, which consists of commands ("jump") and a target file which consists of actions ("I_JUMP"). Given a SCAN split, we take the source file of the training set, and follow a series of steps to obtain partially filled pseudo-constructions.

- **Extracting Candidates:** For every sentence in the train source file, we extract spans of up to length of 4 tokens and add them to the candidate list. We also generate masked spans in which one or more non-consecutive words are replaced by the the token "_", effectively forming a *slot* in a partially filled pseudo-construction. The candidates are then ranked according to their probabilities.

- **Beam Decoding:** We use beam search to segment an input sentence into the best scoring sequence of pseudo-constructions and words. Test source files are not used for mining pseudo-constructions. They are segmented only using the ones induced from the training set.

- **Encouraging Alignment with Target**: Partially filled pseudo-constructions like '_ around _ twice" are advantageous because the same template applies for any fully filled variant - a simple word replacement on the target side works well. However, simple masking also produces "look _ left _" which produces widely different targets for different values of _ and _. Consider,

  look *around* left → I_TURN_LEFT I_LOOK I_TURN_LEFT I_LOOK I_TURN_LEFT I_LOOK I_TURN_LEFT I_LOOK

  look *opposite* left → I_TURN_LEFT I_TURN_LEFT I_LOOK

  To discourage picking candidates like the latter one, we compute an *alignment distance* between the candidate and its equivalent on the target side. For each candidate P, gather the set of source sentences $\mathcal{S}(P) = \{s_1, s_2, \ldots, s_n\}$ in which the pattern occurs, with each source sentence s_i paired to a target sentence t_i. For every $s_i \in \mathcal{S}(P)$, calculate its Levenshtein (edit) distance to every other s_j $(j \neq i)$ in the same set and select the *nearest neighbour*, $\mathrm{NN}(s_i)$—the source sentence that minimises this distance. Let (t_i, t_j) denote the target sentences aligned with $(s_i, \mathrm{NN}(s_i))$.

  Define

  $$\Delta_i = |\mathrm{len}(t_i) - \mathrm{len}(t_j)|$$

  as the absolute difference in their word counts. The resulting *misalignment score* (MS) for pattern P is the average of these differences:

  $$MS(P) = \frac{1}{|\mathcal{S}(P)|} \sum_{i=1}^{|\mathcal{S}(P)|} \Delta_i.$$

  A lower misalignment score indicates that source sequences are more aligned to the target sequences. A pseudo-construction has a low misalignment score when swapping different words into its slots still produces target sentences that look much the same. We add this score as a penalty to the beam search to pick candidates which are more aligned.

Once the source files (train and test) are segmented, we prepare the data for the next stage. For every sentence in the source files, we replace the underscores with slot tokens such as W_n where n refers to the slot number. We save the mapping between the slot tokens and the original words.

The SCAN data consists singleton rules such as *jump* → I_JUMP. We treat this as a bidirectional lexicon. Whenever a token in a target sentence appears in the lexicon, we lookup the source word and then replace it with the associated slot token. For example:

4.2 Training

We use the sequence to sequence transformer architecture as the base model for training purposes, and use the JoeyNMT toolkit (Kreutzer et al., 2019) to train all the models. The model architecture

has an encoder and a decoder each with 4 layers and 4 attention heads with embedding size of 256 and the feed forward layer with the size of 1024. The models are trained for 30 epochs using the NOAM scheduler (Vaswani et al., 2017). Prior to evaluation, we swap back the slot tokens predicted sequence through the mapping saved earlier.

5 Results

The performance on both the splits (ADD JUMP, AROUND RIGHT) is significantly better than the baseline transformer (1) which indicates that we have succeeded in encoding a degree of generalisation through the pseudo-constructions. Overall, they capture reusable structure absent from the flat surface strings, enabling the model to generalise compositionally.

6 Data Efficiency

Compositionality theory posits that exploiting compositional structure enables grasping abstract patterns from far fewer training examples than treating data only at the surface level (Chomsky (1957), Chomsky (1965), Fodor and Pylyshyn (1988)). We test this by training models with smaller samples of the SCAN splits.

After segmentation of the training source file, each sentence is transformed into a series of pseudo-constructions in such a way that multiple sentences might fall into the same resultant sentence type.

look opposite left twice and walk twice → `( W_1 opposite W_2 twice ) and (W_3 thrice)`

jump opposite right twice and run twice → `( W_1 opposite W_2 twice ) and (W_3 thrice)`

To assess the data efficiency of our method we constructed *sentence type–balanced training subsets*, retaining every sentence type but varying the per-type quota $k \in \{1, 3, 5, 10, 25\}$. This produces monotonic subsamples ranging from 5 % to 60 % of the original corpus while guaranteeing full coverage. (Table 1).

On the ADD JUMP split, with only $k = 10$ examples per type—approximately *39 %* of the full training data—the model attains *40.7 %* accuracy, not far from the *47.8 %* trained on entire set.

For the AROUND RIGHT at the same $k = 10$ mark the model reaches merely *9.4 %*, less than half of the *20.4 %* full-data accuracy, and increasing to $k = 25$ (60 %) yields only a marginal gain to *10.2 %*. This pronounced gap reflects the split's higher compositional complexity: mastering the nested "around DIR" construction with repetition operators may require substantially more evidence than the shallow "add jump" pattern.

Split	k / type	Acc. (%)	Size	%
Around Right	1	2.77	741	5
	3	5.98	2,042	13
	5	8.66	3,166	21
	10	9.38	5,423	36
	25	10.18	9,175	60
	Full	20.73	15,225	100
Add Jump	1	12.79	666	5
	3	20.50	1,972	13
	5	29.44	3,178	22
	10	40.69	5,660	39
	Full	47.81	14,670	100

Table 1: Accuracy as the training set is reduced to k examples per type.

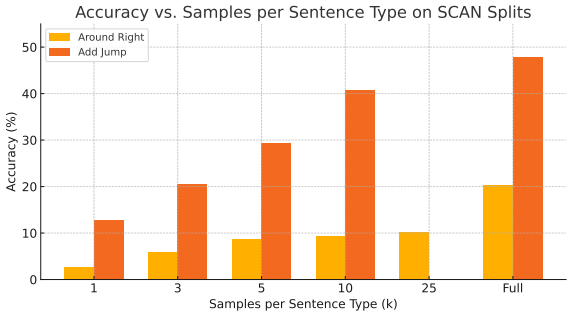


Figure 1: Accuracy versus percentage of full training data for the AROUND RIGHT and ADD JUMP SCAN splits.

The pseudo-construction bias confers strong sample-efficiency benefits on syntactically simple splits (ADD JUMP), but this may not scale to harder generalisation problems (AROUND RIGHT).

7 Conclusion

While we define pseudo-constructions operationally as automatically mined templates, they can be seen as computational approximations to Construction Grammar's notion of conventionalised form–meaning pairings. Unlike fully fossilised or community-shared constructions, pseudo-constructions are data-driven and context-specific, yet they capture structural regularities that support compositional generalisation. Thus, while our primary aim is methodological, the results also lend indirect support to the constructionist hypothesis

that access to reusable schematic patterns is crucial for systematic generalisation. We leave a fuller exploration of their linguistic plausibility and theoretical integration to future work.

A deeper look into errors showed us that our method for finding pseudo-constructions can make several mistakes. For instance, while at first sight "turn around right" and "walk around right" seem to follow the same pattern, their corresponding outputs can vary significantly - this can lead to confusion and failure if the word "turn" is masked away.

We call for more robust approaches into finding constructions in text and for future work into deeper integration of construction processing into neural models.

References

Emre Alp Akyürek and Jacob Andreas. 2022. Lexsym: Discovering and exploiting lexical symmetries for compositional generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2017. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Joost Baan, Dieuwke Hupkes, and Willem Zuidema. 2019. Inspecting the inductive biases of rnns with attentive guidance. In *Proceedings of the 2019 Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford, UK.

Róbert Csordás, Toma Gruber, and Marc Henniges. 2022. Neural data routing: Enforcing compositionality with geometric attention. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

Honghua Dong, Jiayuan Mao, Jiajun Lin, Chuang Wang, Lihong Li, Dengyong Zhou, and Yuncheng Song. 2019. Neural logic machines. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. In Timothy Shopen, editor, *Language Typology and Syntactic Description, Vol. 1*, pages 501–538. Cambridge University Press.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2):3–71.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford, UK.

Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. 2021. Unlocking compositional generalization in pre-trained models using intermediate representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Zihan Huang, Yao Wang, Qian Wu, and Maosong Sun. 2024. Cat: A compositionally aware transformer with multi-primitive decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795. Dataset introduced: *PCFG-SET*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2019. Compositional generalization for neural sequence learning via attentive guidance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhengxuan Jiang and Mohit Bansal. 2021. CGPS-transformer: Compositional generalization by auxiliary sequence prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang,

Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Dataset introduced: *CFQ*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Youngjin Kim, Daniel Keysers, Nathanael Schärli, Daniel Furrer, and Olivier Bousquet. 2021. Structural guidance for transformer self-attention: Hard and soft masking for compositional generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Brenden M. Lake. 2019. Compositional generalization through meta sequence-to-sequence learning. In *Advances in Neural Information Processing Systems (NeurIPS) 32*.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: Strong tests for compositionality in humans and machines. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Dataset introduced: SCAN.

Brenden M. Lake and Marco Baroni. 2023. Meta-in-context learning induces human-like compositional generalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar, Volume 1: Theoretical Prerequisites*. Stanford University Press, Stanford, CA.

Jialin Li, Shuwen Lu, Yang Liu, and Mohit Bansal. 2022. Neural-symbolic recursive machines for systematic generalization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ruixiang Li, Yichao Chen, Sheng Lin, and Marco Baroni. 2019. CGPS-rss: Separating primitive and functional channels improves compositional generalization. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS) — Compositionality Workshop*.

X. Li and colleagues. 2021. Cognition: A dataset for testing compositional generalisation in neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages xx–yy, Online. Association for Computational Linguistics. Dataset introduced: *COGNITION*.

Adam Liška, Germán Kruszewski, and Marco Baroni. 2018. Memorize or generalize? searching for a compositional RNN in a haystack. *arXiv preprint arXiv:1802.06467*.

Cang Liu, Pengcheng Zhou, Zhenzhong Lan, and Yang Wang. 2022. Lane: Learning analytical expressions for systematic generalization. In *Advances in Neural Information Processing Systems (NeurIPS) 35*.

João Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium. Association for Computational Linguistics.

Gary F. Marcus. 2003. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press, Cambridge, MA.

Maxwell Nye, Armando Solar-Lezama, and Joshua B. Tenenbaum. 2020. Compositional generalization by program synthesis. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Barbara H. Partee, Alice ter Meulen, and Robert E. Wall. 1990. Compositionality. In *Mathematical Methods in Linguistics*, pages 319–334. Springer.

Chen Qiu, Yutong Yu, Emre Alp Akyürek, and Jacob Andreas. 2022. Csl: Compositional structure learner for data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jacob Russin, Jianshu Jo, and Randall C. O'Reilly. 2019. Compositional generalization in sequence-to-sequence models via syntactic attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the*

28th International Conference on Computational Linguistics, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pages 5998–6008.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Shuyan Zhu, Zhengxuan Jiang, Ruixiang Li, and Mohit Bansal. 2021. DUEL: Transferable compositional inductive bias via meta-learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

From Form to Function: A Constructional NLI Benchmark

Claire Bonial[1], **Taylor Pellegrin**[2], **Melissa Torgbi**[3], **Harish Tayyar Madabushi**[3],
[1]DEVCOM Army Research Laboratory, U.S.A.
[2]Oak Ridge Associated Universities, U.S.A. [3]University of Bath, U.K.
claire.n.bonial.civ@army.mil

Abstract

We present CoGS-NLI, a Natural Language Inference (NLI) evaluation benchmark testing understanding of English phrasal constructions drawn from the Construction Grammar Schematicity (CoGS) corpus. This dataset of 1,500 NLI triples facilitates assessment of constructional understanding in a downstream inference task. We present an evaluation benchmark based on the performance of two language models, where we vary the number and kinds of examples given in the prompt, with and without chain-of-thought prompting. The best-performing model and prompt combination achieves a strong overall accuracy of .94 when provided in-context learning examples with the target phrasal constructions, whereas providing additional general NLI examples hurts performance. This evidences the value of resources explicitly capturing the semantics of phrasal constructions, while our qualitative analysis suggests caveats in assuming this performance indicates a deep understanding of constructional semantics.

1 Introduction

This research addresses the challenge of how we determine what computational systems know of a language; specifically, we focus on the large portion of the English language in which meaning goes beyond the sum of lexical parts—phrasal constructions. Whereas our past NLP tools were developed and therefore grounded in some form of grammatical theory (e.g., phrase structure or dependency parsing), LLMs lack grounding in linguistic theory. Instead, their development is based on the encoder-decoder architecture, which was originally designed for sequence-to-sequence tasks, specifically translation (Bahdanau et al., 2016). This dichotomy impedes methods for evaluating LLMs, as their performance on meta-linguistic tasks, such as semantic role labeling, which previously served as benchmarks for the individual components in an NLP pipeline, are poor predictors of LLM fluency on downstream applications.

Premise	I had brushed my hair smooth.
Hypothesis	I had smooth hair because I brushed it.
Relation	Entailment

Table 1: CoGS-NLI example for a premise including the Resultative cxn; inferring the entailment relies upon recognition of the constructinoal semantics.

Although LLMs lack theoretical grounding, evaluation of language proficiency benefits from analysis through a particular theoretical lens, which enables one to hypothesize the appropriate formal units of a language and the way in which meaning is associated with those formal units. We leverage Construction Grammar (CxG) to analyze language (specifically English) as a set of constructions (cxn), or pairings of meaning and form at any structural level, including morphemes, lexemes, and phrases. As a usage-based linguistic theory, CxG provides an experimentally-validated framework for how speakers acquire language and generalize knowledge of frequently heard cxns to totally creative and novel instantiations (e.g., Tomasello (2009); Johnson and Goldberg (2013)). **CxG research demonstrates that speakers attribute meaning to special syntactic templates (phrasal cxns)—meaning that goes beyond that of the individual lexical items alone; CoGS-NLI allows us to evaluate if LLMs also attribute the appropriate meaning to phrasal cxns.**

We leverage Construction Grammar Schematicity (CoGS) corpus instances (Section 2) as the premises in the subsequent development of a comprehensive dataset of 1500 Natural Language Inference (NLI) triples (see Table 1), which serves as a downstream test of functional understanding of cxns (Section 3). We benchmark performance on

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 172–179

this task with two models (GPT-3.5-turbo, GPT-4o), and demonstrate that including examples with constructional premises in few-shot prompting boosts performance to reach a top-end accuracy of .94 (Section 4).[1] This shows first that resources exemplifying the target constructional semantics are beneficial to performance, and second that constructional premises do not pose a problem for state-of-the-art models in this task. However, there is qualitative evidence that tempers the conclusion that models must grasp constructional semantics in order to perform successfully on the task (Section 5). We close with recommendations for future steps in evaluating constructional understanding (Section 6).

2 Related Work

Related work in the area of evaluating LLMs through the lens of CxG fall broadly into two types of research: i. testing for LLM recognition and classification of certain cxns; and ii. testing for LLM functional understanding

In the first area, Tayyar Madabushi et al. (2020) demonstrated that a variety of base and fine-tuned BERT models are able to distinguish between sentences that instantiate a particular cxn and those that do not. Li et al. (2019) recreate a psycholinguistic test in which models of varying sizes are tested for their ability to group sentences by semantic similarity, where some sentences include the same cxn (e.g., Caused-motion), and others involve different cxns but semantically similar lexical verbs (e.g., sneeze, burp). The authors find that while the smallest language model with 1 million parameters, MiniBERTas (Pérez-Mayos et al., 2021), groups the sentences according to lexical semantics, the largest model with 30 billion parameters, RoBERTa (Liu et al., 2019), groups sentences according to constructional semantics. Of particular relevance to this research, Bonial and Tayyar Madabushi (2024a) develop the initial test set of corpus examples of cxns later released as the CoGS corpus, and test larger models (GPT-3 and 4) for recognition of sentences containing a cxn. The authors find a clear trend demonstrating that the models can recognize substantive cxns with some fixed words (e.g., Much-less), but have increasing difficulty recognizing cxns of increasing schematicity or variability.

[1]The evaluation data subset, prompts, and outputs can be found here: https://github.com/melissatorgbi/from-form-to-function

Overall, the research in the first area demonstrates that while models can recognize and classify some cxns, more abstract cxns present a problem for recognition. Furthermore, studies of recognition and classification do not directly demonstrate whether or not LLMs are proficient users of the cxns of a language; i.e. whether or not the models "understand" the constructional semantics.

Thus, we emphasize the importance of the second area of research, which aims to test LLM functional understanding of cxns in a downstream task. Both Weissweiler et al. (2022) and Zhou et al. (2024) set up evaluations of formal recognition of cxns as well as semantic understanding of the Comparative-correlative and Causal-excess cxns respectively. In both cases, the authors find that models are able to distinguish the cxns, but perform poorly on tests of semantic understanding in the form of downstream questions. Similarly, Scivetti et al. (2025a) finds that smaller-scale LLMs are sensitive to the formal properties of the Let-alone cxn, but reflect no sensitivity to the semantic properties, again in a set of downstream questions testing for understanding.

3 Dataset Development

NLI is a task in which a premise is presented followed by a hypothesis, and the task is to determine if the hypothesis i. must be true given the premise (entailed); ii. may or may not be true given the premise (neutral); iii. cannot be true given the premise (contradicted). We base our task guidelines on the Stanford NLI (SNLI) corpus, which was developed to test semantic representations, as the authors consider understanding entailment and contradiction to be fundamental to natural language understanding (Bowman et al., 2015). NLI has since been adopted as a relatively common test of semantic understanding with several community evaluations (e.g., Marelli et al. (2014); Lee et al. (2024)). As a result, there is widespread availability of NLI data on the web, and it is a relatively common benchmark for LLMs. This also influenced our choice—as there is abundant data on LLM performance for the NLI task, we can distinguish baseline abilities of models on this task from performance on the constructional variant (Sarlin et al., 2020; Raffel et al., 2020; Wei et al., 2022).

We draw our premises from the corpus instances of the 10 cxn types in CoGS (Bonial and Tayyar Madabushi, 2024b); there are about 50 unique corpus instances of each cxn type, giving us about

500 unique premises. The cxns in CoGS vary in *schematicity* (how many words of constructional slots are substantive/fixed or schematic/variable), which enables us to test constructional understanding for fixed-word cxns in which meaning is consistently associated with a particular form, as well as variable-word cxns, in which meaning is associated with templatic syntactic patterns (such as the DITRANSITIVE: The student [noun phrase] handed [verb] the teacher [noun phrase] a book [noun phrase]—i.e., NP V NP NP). A listing of all cxns and example NLI triples from CoGS-NLI is given in Appendix B, Table 4.

One native English speaker (and author of this paper) with an undergraduate degree in Linguistics (but no training in CxG specifically) was given a spreadsheet of the CoGS premises and asked to produce 3 NLI triples—an entailed, neutral, and contradicted hypothesis for each premise; thus, the corpus totals 1500 triples associated with 500 unique premises. We provide guidelines adapted from SNLI definitions of the relations. The NLI author selected triples to create in any order desired to prevent getting stuck on more difficult cases. Depending on the length and complexity of the premise, the hypotheses could take several minutes to process, or come to the author immediately. Overall, the development of the CoGS-NLI corpus was done over the course of a year to prevent fatigue and degraded quality.

We conducted several quality checks of the CoGS-NLI corpus by comparing agreement on the assigned relation of subsets of data (totaling 441 NLI instances) across three annotators (and authors of this paper) against the author's originally assigned relation. Percentage agreement on the initial set of triples ranged from 71-80%, or .55-.70 when measured as Cohen's κ, indicating substantial agreement. All disagreements were revisited, and a second author reworded the hypotheses. Agreement on the reworded hypotheses then reached 89%, or .84 Cohen's κ, indicating very strong agreement equal to the published agreement of individual annotators with respect to gold relation for SNLI.

4 Evaluation Experiments

4.1 Methodology

We provide a performance benchmark by testing models on the same subset of the data that was evaluated for human agreement. Specifically, we hold out 50 instances for in-context learning and use the remaining 391 instances as the test set. The in-context learning examples were randomly chosen where each example contains a single premise with a neutral hypothesis, entailment hypothesis and contradiction hypothesis. The in-context learning examples provided are paired with target phrasal cxns in the test set in order to provide clear examples of the phrasal constructional semantics within the NLI task.

Setting	IC Data	GPT-3.5	GPT-4o
0-shot	None	0.74	0.89
1-shot	CoGS-NLI	0.78	0.91
3-shot	CoGS-NLI	0.83	**0.94**
1-shot	SNLI	0.70	0.89
3-shot	SNLI	0.69	0.90

Table 2: Evaluation results, reported in accuracy, on the CoGS-NLI dataset. "IC Data" refers to the type of data used as in-context examples.

We evaluate GPT-4o-2024-05-13 and GPT-3.5-turbo-0125 models; these models were chosen as representatives of LLM capabilities due to their large size. The temperature is set to 0 to minimize randomness in the model outputs.

We compared results for six different prompt variations, with and without explicitly prompting for Chain of Thought (CoT). We report results for our best-performing prompt, provided in full in Appendix A. We also experimented with 0-shot through 3-shot learning, with two different sources of examples: held-out examples from the CoGS-NLI dataset and selected examples with full-sentence premises from the SNLI corpus. We conduct this comparison in order to determine if the constructional examples boost performance, or if general SNLI examples are sufficient. Note that the CoGS-NLI examples include the target phrasal cxns included in the evaluation, providing clear examples of how these cxns should be interpreted with respect to the NLI task. While the SNLI examples also include cxns of English, they do not include the target phrasal cxns of CoGS.

4.2 Results

Results are reported in Table 2. We see a 5-point boost in performance in the 3-shot setting with constructional examples and achieve a top-end performance of 94% accuracy from GPT-4o. We do not see an equivalent boost in GPT-4o performance in the 3-shot setting with general SNLI examples. The constructional examples are even more helpful for GPT-3.5, where 3-shot outperforms zero-shot

Premise 1	Constance squeezed her way down the platform looking for the first-class carriages.
Hypothesis	Constance waited in line for the first-class carriages.
Relation	Gold: Contradiction; GPT-4o: Neutral
Premise 2	The 23 frantically scrambled to the rear of the sub.
Hypothesis	The 23 were calm at the rear of the sub.
Relation	Gold: Contradiction; GPT-4o: Contradiction

Table 3: Premise 1 exemplifies an error for the most frequently mis-analyzed cxn (Way-manner). Premise 2 (Intransitive-motion) exemplifies a hypothesis with information outside of the constructional semantics that cues the contradiction (i.e. "frantically" vs. "calm".)

by 9 points. Notably, the 3-shot setting with SNLI examples actually *hurts* performance by 5 points.

5 Discussion

Given that the CoGS developers found that models were able to recognize and classify substantive cxns (with fixed words) with much greater accuracy than schematic cxns (with no fixed words and only variable syntactic-semantic slots) (Bonial and Tayyar Madabushi, 2024a), we also assessed if there were performance differences in CoGS-NLI for those cxns classed as fully fixed/substantive, partially fixed, or fully variable/schematic. In contrast to the earlier findings, we do not find a notable difference in performance based upon the schematicity level of the cxn in the premise (see Appendix B Table 5 for performance results separated by phrasal cxn type). However, when we analyze distinct cxns, GPT-4o achieves the highest accuracy on the fully variable Resultative cxn (see Table 1) and the lowest accuracy on the partially variable Way-manner cxn. We provide an error case in (Premise 1) of Table 3. The stronger performance that we see on schematic cxns like the Resultative in the functional understanding NLI task may relate to the frequency of the cxn—LLMs may be better at "understanding" more frequent cxns with greater representation in pretraining data, and the fully schematic argument structure cxns of CoGS are also some of the most frequent cxns of English. We begin to explore this question further in ongoing research (Scivetti et al., 2025b).

The performance of both models on the CoGS-NLI dataset is comparable to performance on SNLI (Ye et al., 2023; OpenAI et al., 2024); thus, we can conclude that including constructional premises does not pose a significant challenge in this task. We note two intertwined limitations in drawing the strong conclusion that these models therefore have a functional understanding of the semantics of the cxn. First, in the NLI task generally, models may rely on spurious features (e.g., the number of tokens) of the premise and hypothesis to solve the task without actually understanding the constructional semantics (Gururangan et al., 2018). Second, the hypotheses may probe other aspects of meaning of the premise outside of the constructional semantics. Premise 2 in Table 3 provides an example where the hypothesis includes the modifier "calm" which contradicts the modifier "frantically" in the premise, but bears no relation to an understanding of the Intransitive-motion constructional semantics. Taken together, these limitations mean that NLI task solvability generally, including that of CoGS-NLI, may be correlated with features outside of a deep semantic understanding.

Thus, on the whole, our results demonstrate that while constructional resources are needed for boosting performance on downstream tasks in which language includes phrasal cxns (note that this is not rare—argument structure cxns are some of the most common phrasal cxns of English), more precise probing evaluations are needed for assessing constructional understanding.

We take steps to craft hypotheses that more precisely target the constructional semantics in Scivetti et al. (2025b). This research also leverages a subset of the CoGS corpus for premises in setting up an NLI evaluation of constructional understanding; however, unlike the present research, we semi-automatically generate the NLI triples by leveraging templates for the neutral, contradicted and entailed hypotheses across all instances of a given cxn type. We note that the templatically generated NLI triples may inadvertently simplify the task by consistently patterning different types of hypotheses. In contrast, CoGS-NLI enables testing understanding through NLI while leveraging free-form, human-authored triples. Together, CoGS-NLI and the templatically-generated NLI dataset of Scivetti et al. (2025b) provide complementary evaluation resources.

6 Conclusions & Future Work

The evaluation of constructional information encoded in LLMs has been approached in several ways. A significant limitation of early methods, such as probing internal model weights, is that discovering the presence of constructional information encoded in weights does not guarantee it is functionally utilized. While prompting allows us to observe how models interact with cxns, metalinguistic tasks that test an LLM's ability to identify sentences as instances of the same cxn measure a classificatory skill, not whether the model can make use of that cxn's meaning to solve a problem.

Our work directly addresses this gap by focusing on the functional application of constructional knowledge. To this end, we created an NLI dataset where premises are carefully selected to feature specific cxns. Our results show that including examples with constructional premises does boost performance, indicating a value to constructional resources like CoGS-NLI. While our results suggest that current models can often correctly solve this task, we recognize that the NLI task does not always isolate the exact semantic meaning carried by the cxn itself. Therefore, in our ongoing and future work we are developing more targeted evaluations to verify that an LLM's reasoning is guided by the precise meaning conveyed by a grammatical cxn (Scivetti et al., 2025b).

Furthermore, any claim about an LLM's *understanding* must contend with recent findings that their performance relies on "context-directed extrapolating from training data priors" (Tayyar Madabushi et al., 2025). Therefore, to genuinely test a model's reasoning capabilities, it is not enough to evaluate it on problems for which priors readily exist in model training data. A systematic evaluation must present novel scenarios with minimal or nonexistent priors, forcing the model to demonstrate *inherent* 'reasoning' or 'understanding' rather than relying on statistical shortcuts. We will continue to leverage CxG as a formalism for targeting language that is creative and novel, but readily understandable by people in order to support such systematic evaluation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Claire Bonial and Harish Tayyar Madabushi. 2024a. Constructing Understanding: on the Constructional Information Encoded in Large Language Models. *Language Resources and Evaluation*, pages 1–40.

Claire Bonial and Harish Tayyar Madabushi. 2024b. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Lung-Hao Lee, Chen-Ya Chiou, and Tzu-Mi Lin. 2024. Nycu-nlp at semeval-2024 task 2: Aggregating large language models in biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1455–1462.

Hao Li, Wei Lu, Pengjun Xie, and Linlin Li. 2019. Neural Chinese address parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3421–3431, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. 2024. Gpt-4o system card.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4937–4946, Los Alamitos, CA, USA. IEEE Computer Society.

Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025a. Unpacking let alone: Human-scale models generalize to a rare construction in form but not meaning. *arXiv preprint arXiv:2506.04408*.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025b. Assessing language comprehension in large language models using construction grammar.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Harish Tayyar Madabushi, Melissa Torgbi, and Claire Bonial. 2025. Neither stochastic parroting nor agi: Llms solve tasks through context-directed extrapolation from training data priors.

Michael Tomasello. 2009. The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 3804–3811, Torino, Italia. ELRA and ICCL.

A Prompts

The following prompt was the best performing variation, achieving .94 accuracy with gpt-4o.

Prompt 1:

"You are the world's best annotator. You are tasked with annotating a triple for Natural Language Inference. You must determine the inference relation between the Premise and the Hypothesis by selecting one of three numerical codes that reflect the relationship:
0 – Entailment: The Hypothesis is definitely true given the Premise.
1 – Neutral: The Hypothesis may or may not be true given the Premise.
2 – Contradiction: The Hypothesis cannot be true given the Premise.
Output a single numerical value between 0 and 2 inclusive, corresponding to the associated relation."

B CoGS-NLI Constructions: Examples & Results by Construction

We provide a listing of all ten cxns included in CoGS-NLI, along with example NLI triples, in Table 4. We then provide performance results across individual cxn types in Table 5.

Construction	Premise, Hypothesis	Relation
Much-less	P: When my dad catches swarms sometimes he doesn't even wear a veil, much less a bee suit. H: When my dad handles swarms, he sometimes wears a veil.	Entailment
Let-alone	P: None of these arguments is notably strong, let alone conclusive. H: All of the given arguments are strong and conclusive.	Contradiction
Way-manner	P: As she felt her way forward, suddenly a knight on horseback galloped past her. H: She was moving forward when a knight on horseback almost ran her over.	Neutral
Comparative-correlative	P: The fewer things we make the more sustainable we are. H: We are more sustainable if we make fewer things.	Entailment
Causative-with	P: The waiter filled her glass with white wine. H: She ordered the white wine in a glass.	Neutral
Conative	P: He nibbled at the filet, then ate ravenously. H: He took big bites of the filet, then slowed down.	Contradiction
Ditransitive	P: They threw me a surprise party. H: They forgot to give me a surprise party.	Contradiction
Caused-motion	P: The MiG-25 fired an AAM at the Predator. H: The MiG-25 tried to hit the Predator.	Neutral
Intransitive-motion	P: Armed troops marched to the substations and turned the power back on. H: The power was turned back on by armed troops that marched to the substations.	Entailment
Resultative	P: He ate himself sick. H: He felt wick from eating.	Entailment

Table 4: One example for each of the ten phrasal cxns included in CoGS-NLI. Note that premises are drawn directly from CoGS, and CoGS-NLI contributes three hypotheses for each premise: one entailed, one contradicted, and one neutral.

Setting	IC Data	Construction	GPT-3.5	GPT-4o
0-shot	None	Let-alone	0.79	0.92
		Way-manner	0.58	0.79
		Comparative- correlative	0.60	0.70
		Causative- with	0.83	0.94
		Conative	0.69	0.88
		Caused- motion	0.78	0.92
		Intransitive- motion	0.78	0.91
		Resultative	0.80	0.94
1-shot	CoGS-NLI	Let-alone	0.83	0.92
		Way-manner	0.64	0.79
		Comparative- correlative	0.57	0.73
		Causative- with	0.85	0.93
		Conative	0.88	0.91
		Caused- motion	0.81	0.97
		Intransitive- motion	0.74	0.94
		Resultative	0.79	0.94
3-shot	CoGS-NLI	Let-alone	0.67	0.92
		Way-manner	0.79	0.85
		Comparative- correlative	0.67	0.87
		Causative- with	0.89	0.94
		Conative	0.90	0.94
		Caused- motion	0.83	0.97
		Intransitive- motion	0.86	0.97
		Resultative	0.80	0.98
1-shot	SNLI	Let-alone	0.79	0.92
		Way-manner	0.52	0.88
		Comparative- correlative	0.60	0.73
		Causative- with	0.78	0.93
		Conative	0.69	0.86
		Caused- motion	0.69	0.92
		Intransitive- motion	0.78	0.93
		Resultative	0.67	0.91
3-shot	SNLI	Let-alone	0.62	0.92
		Way-manner	0.58	0.85
		Comparative- correlative	0.63	0.67
		Causative- with	0.78	0.93
		Conative	0.71	0.91
		Caused- motion	0.67	0.92
		Intransitive- motion	0.72	0.96
		Resultative	0.68	0.92

Table 5: Evaluation results, reported in accuracy, on the CoGS-NLI dataset for the best performing prompt for each individual construction. “IC Data" refers to the type of data used as in-context examples.

Evaluating CxG Generalisation in LLMs via Construction-Based NLI Fine Tuning

Tom Mackintosh[1], **Harish Tayyar Madabushi**[1], **Claire Bonial**[2],
[1]University of Bath, U.K.
[2]DEVCOM Army Research Laboratory, U.S.A.
htm43@bath.ac.uk

Abstract

We probe large language models' ability to learn deep form-meaning mappings as defined by construction grammars. We introduce the ConTest-NLI benchmark of 80k sentences covering eight English constructions from highly lexicalized to highly schematic. Our pipeline generates diverse synthetic NLI triples via templating and the application of a model-in-the-loop filter. This provides aspects of human validation to ensure challenge and label reliability. Zero-shot tests on leading LLMs reveal a 24% drop in accuracy between naturalistic (88%) and adversarial data (64%), with schematic patterns proving hardest. Fine-tuning on a subset of ConTest-NLI yields up to 9% improvement, yet our results highlight persistent abstraction gaps in current LLMs and offer a scalable framework for evaluating construction-informed learning.

Model	Constr. Semantics	Constr. Distinction
Prior work (Scivetti et al., 2025)		
GPT-4o	0.88	0.58
GPT-o1	0.90	0.46
Llama 3 70B	0.74	0.52
Human	0.90	0.83
This work		
Llama-3.1-8B (baseline)	0.57	0.33
Llama-3.1-8B (fine-tuned)	0.66	0.39

Table 1: Comparison of model performance on constructional (constr.) understanding. The top section, with results from prior work Scivetti et al. (2025), shows that LLMs struggle with the constructional distinction task compared to the human baseline. The bottom section presents our results, showing that this shortcoming persists despite fine-tuning. See Section 6 for full results.

1 Introduction and Motivation

Human intelligence is often attributed to our capacity for language — and, in particular, our ability to generalize abstract, compositional meaning from surface structure (Pinker, 2003). Construction Grammar (CxG) (Goldberg, 1995; Croft, 2001; Tayyar Madabushi et al., 2020) (See also Section 2) formalises this by treating linguistic knowledge as form-meaning pairings — constructions — that range from single words to complex syntactic frames. Understanding whether large language models (LLMs) acquire such abstractions remains a fundamental question at the intersection of linguistics and artificial intelligence.

In CxG, each construction pairs a conventionalised *form* with an associated meaning. The form is the syntactic configuration, possibly including fixed lexical items, while the meaning is provided by the construction as a whole rather than from the individual lexical items. For example, the Resultative construction has the form Noun Phrase (NP), Verb (V), Noun Phrase (NP), Adjective (ADJ) and the meaning "the action described by the verb causes the object to enter the state described by the adjective" (Goldberg, 1992). In "She hammered the metal flat," the state 'flat' is the *result* of the hammering event, a meaning supplied by the Resultative.

While each construction has a specific *form*, different constructions can share the same syntactic structure. For instance, the Depictive construction also uses the NP V NP ADJ *form* but has a distinct *meaning* (Goldberg and Jackendoff, 2004). In the Depictive, the adjective describes the state of the noun *during* the action of the verb, not as a result of it. This is illustrated by the example, "A famous emperor buried scholars alive." Here, 'alive' describes the state of the scholars while they were being buried; crucially, the act of burying did not *cause*

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 180–189

Construction	Premise	Hypothesis	Label
Resultative	Through effort, the gardener cultivated the garden lush.	The gardener worked hard to create a vibrant outdoor space.	Entailment
Caused Motion	The magician levitated the rabbit into the hat.	The magician placed the rabbit on the table.	Contradiction
Causative With	In no time, the magician had filled the auditorium with applause.	The magician performed in different auditoriums.	Neutral

Table 2: Examples drawn from the ConTest-NLI training set, with one instance of each NLI label from distinct constructions.

them to become alive. This distinction highlights how syntactically identical sentences can convey vastly different meanings based on the underlying construction.

Recent evaluation work (Scivetti et al., 2025), which used the downstream task of Natural Language Inference (NLI) to create a test of the functional understanding of LLMs, reveals that while LLMs can correctly interpret an entrenched construction like the Resultative even with unusual lexical items, their generalization ability is limited. Specifically, when presented with creative instances of a less entrenched construction like the Depictive, LLMs tend to overgeneralize and assign the meaning of the more frequent, or entrenched, Resultative construction. Indeed they overgeneralise to such an extend that they show a performance drop of over 40% on on this task, when compared to the original task of interprating the meaning of entrenched constructions. These findings are summarized in Table 1. This failure to use lexical and pragmatic cues to resolve syntactic ambiguity, a task at which native speakers can perform quite easily, demonstrates that the models' grasp of abstract meaning remains brittle and overly dependent on statistical patterns rather than a robust, human-like linguistic competence.

While Scivetti et al. (2025) identify this shortcoming, they leave the reasons for this specific failure to future work. Therefore, this work aims to answer this question by examining the model's expertise with the task. Specifically, **we hypothesize and investigate if training the model on explicit NLI examples will help the model better 'understand' creative, less entrenched constructions in the presence of a more frequent distractor.** A positive result would offer a clear-cut path to improving these models' understanding, whereas a negative result would point to a more fundamental issue that needs to be addressed.

To this end, this paper introduces **ConTest-NLI** (Constructional Test Natural Language Inference), a scalable dataset designed to evaluate whether LLMs internalize the semantics of linguistic constructions or rely solely on surface heuristics. ConTest-NLI specifically targets systematic generalization across unseen verbs, arguments, and constructions. This provides a scalable way to inform LLMs of specific construction examples, allowing a new control for deeper research into semantic understanding of linguistic theory.

One key empirical finding is that **LLMs fail to generalize constructional semantics across syntactically identical but semantically distinct constructions**. For example, models trained to detect entailment violations in the Resultative construction show no improved performance on the Depictive construction, despite their shared syntax. This lack of transfer reveals that current models do not acquire construction-general semantics, but instead overfit to narrow instantiations.

To test our hypothesis at scale, we use a semi-automated pipeline that facilitates generation of synthetic constructional NLI triples: ConTest-NLI. Example data is shown in Table 2. Our pipeline leverages syntax-informed template generation of eight core constructions and model-in-the-loop filtering to identify deceptive false positives.

We compare ConTest-NLI to two existing CxG benchmarks from Scivetti et al. (2025): the manually curated Construction-NLI (CxN-NLI), and the more challenging Construction-NLI-Distinction (CxN-NLI-Distinction), which introduces false positives that share syntax but diverge in semantics. While those datasets offer excellent linguistic control, they remain small and difficult to scale. ConTest-NLI complements them by enabling controlled experiments across a broader constructional space, yielding more robust insights into model generalization.

We fine-tune small-scale LLMs (LLaMA 3.1 8B Instruct, Mistral 8B Instruct) on ConTest-NLI examples and evaluate their performance across both

seen and unseen constructions. While models improve (≤9pp) on the trained construction, their failure to generalise — especially to constructions with shared syntactic structure — suggests a fundamental limitation in semantic abstraction.

ConTest-NLI is thus shown to be useful for evaluating systematic language understanding in LLMs, bridging the scale of automated generation with the precision of theoretical linguistics. In our experiment, we use ConTest-NLI to gather direct empirical evidence that shows, without further architectural or training innovations, LLMs do not acquire transferable constructional semantics — highlighting a key divergence from human-like generalization.

2 Related Work

CxG is a linguistic theory that positions constructions — form-meaning pairings — as the fundamental units of language. A construction, as defined within this framework, is any linguistic pattern whose meaning is not fully predictable from its individual components (Goldberg, 1995; Croft, 2001; Tayyar Madabushi et al., 2020).

Further cognitive and usage-based studies within CxG emphasize that humans generalize constructional meanings from frequency of exposure and exemplar experiences. Psycholinguistic research, notably by Bencini and Goldberg (2000), showed that participants' interpretations of sentence meanings significantly reflect constructional semantics rather than just verb meanings alone. In their experiment, participants grouped sentences primarily by the underlying constructional meaning, demonstrating that constructions themselves carry cognitive reality independent of specific lexical content (Kaschak, 2007; Goldberg et al., 2007).

This perspective is particularly relevant for evaluating language comprehension in computational models. Recent computational linguistic research leverages CxG to systematically assess language understanding in large language models. Studies such as those by Tayyar Madabushi et al. (2020) and Scivetti et al. (2025) illustrate how CxG provides a robust theoretical grounding to create targeted, semantically-rich evaluations for LLMs. These studies specifically demonstrate the utility of construction-based Natural Language Inference (NLI) tasks, highlighting significant limitations of LLMs in generalizing abstract constructional semantics when faced with novel linguistic contexts or minimally represented constructions (Bonial and Tayyar Madabushi, 2024).

Thus, CxG not only provides insights into human linguistic competence but also offers a rigorous toolset for probing and understanding the boundaries of true semantic generalization in language models — a foundational concern of contemporary NLP research.

Also, constructional semantics provide a controlled yet diverse linguistic testbed. Constructions vary significantly in their schematicity — from highly substantive, lexically fixed forms, such as the Let-alone construction, to more abstract and schematic patterns such as Resultative or Caused-motion constructions (Bonial and Tayyar Madabushi, 2024; Scivetti et al., 2025). Evaluations across this spectrum enable systematic testing of LLMs' capacity for abstract semantic generalization. Crucially, previous computational studies demonstrate that while LLMs may perform well on lexically anchored constructions due to frequency and memorization, their performance substantially deteriorates when faced with more schematic and less frequent constructions (Weissweiler et al., 2022; Scivetti et al., 2025).

Despite the promise of construction grammars as a diagnostic for true semantic generalization, existing computational CxG evaluations remain narrowly focused, limited in scale, or insufficiently controlled. This project fills that gap by introducing a semi-automated pipeline that combines high-variance templating to produce large-scale CxG evaluation data.

3 ConTest-NLI Dataset Development

ConTest-NLI is designed as a scalable, high-variance training resource for probing whether LLMs can learn and generalise the semantics of English constructions beyond rote lexical recall. It forms the centrepiece of a broader multi-corpus strategy, enabling both in-domain fine-tuning and rigorous, out-of-distribution testing. This is essential: single-source datasets are prone to heuristic exploitation, whereas orthogonal axes — synthetic vs. natural, fluent vs. adversarial — allow us to pinpoint exactly where generalisation succeeds or fails.

We adopt the eight English constructions from Scivetti et al. (2025), spanning the substantive–schematic continuum (e.g., Let-alone vs. Resultative). Construction details and examples are pro-

vided in Appendix A. Each construction is instantiated by ≥10,000 examples, generated from ≥8 canonical templates varying surface order, clause type, and optional modifiers.

3.1 Template Engineering

For each construction, we hand-crafted 8–12 canonical skeletal templates encoding the obligatory syntactic positions and any construction-specific function words (e.g., "The more X, the more Y" for the Comparative Correlative). These templates are designed to maximise *controlled diversity*: varying word order, clause type, voice (active/passive), adjunct position, and optional modifiers ensures that no single surface pattern dominates.

Lexical slots are populated from "mid-frequency" lemmas (20–60th percentile in BookCorpus) to reduce overlap with model pretraining data. We further expand these lists using WordNet synonyms, hyponyms, and antonyms, while explicitly excluding the top 10,000 Common Crawl tokens and any lexemes whose semantics would trivially satisfy the inference (e.g., *moved* in a Caused-Motion frame). Controlled adverbial pools (manner, time, frequency, intensity) and automated morphological inflection via `lemminflect` add stylistic variation without altering truth-conditional content. Examples of templates and their instantiations are provided in Table 3, and templates for all constructions are provided in Appendix B.

Construction	Example Template / Instantiation
RESULTATIVE	"The [agent] [verb] the [patient] [end-state]" → *The chef chopped the carrots thin*
CAUSED-MOTION	"X [verb] Y into Z" → *They rolled the log into the river*
CAUSATIVE-WITH	"X filled C with S" → *The artist filled the gallery with vibrant paintings*
LET-ALONE	"Even getting X to [verb] was tough, let alone Y" → *Even getting the robot to succeed was tough, let alone the knapsack*

Table 3: Sample templates and instantiations from the ConTest-NLI generation pipeline. Note that template filling results in some semantic infelicity, such as the Let-alone comparison of a robot and a knapsack.

3.2 Example Generation

Each premise sentence is paired with three hypotheses labelled *entailment* (E), *neutral* (N), or *contradiction* (C), with labels assigned via construction-specific generation rules grounded in formal semantics. For example, a CAUSATIVE-WITH premise *The artist filled the gallery with vibrant paintings* yields:

- (E) *The gallery contained vibrant paintings*
- (C) *The gallery was empty of any paintings*
- (N) *The artist painted in a nearby studio*

This approach ensures that all examples are fluent and natural-sounding, while still requiring the model to attend to the construction's form-meaning pairing to make the correct inference.

3.3 Manual Analysis

We conducted manual analysis to ensure the quality of the dataset along two dimensions: (i) the generated constructions are indeed members of the specified construction type, and (ii) the established relation for each NLI triple is accurate. We randomly sampled 100 instances of our dataset, balanced across neutral, entailment, and contradiction relations. One author and native English speaker, trained in linguistics and CxG, provided a binary rating for (i) and (ii), and where the author disagreed with the relation provided, gave a corrected NLI relation. The result of this analysis was that 99/100 instances were judged to be instances of the specified construction type, and 94/100 NLI instances were judged to have the correct relation. This indicates the overall high quality of the developed dataset.

However, the manual analysis further revealed two limitations of the synthetic NLI triples. First, the data in the sample were relatively repetitive. While we expect repetitions of the premise with unique hypotheses representing different entailment relations to the premise, we found that the hypotheses themselves were also somewhat repetitive, sometimes differing only in a single word (e.g., "tree trunk" vs. "tree bark" or adding "might"). Second, judging the entailment relation was somewhat trivial for many triples, given that entailed hypotheses were sometimes near-verbatim repetitions of the premise, whereas contradicted hypotheses often leveraged a single lexical item of opposite semantics to a counterpart in the premise. We note that manual development of NLI triples can also lead to the same limitations.

3.4 Dataset Splits

We enforce a deterministic 70/15/15 train/dev/test split *within each construction*. Crucially, the split is

lexeme-held-out: any verb lemma appearing in the test set for a given construction is entirely absent from its train and dev sets. This protocol is applied consistently across all ConTest-NLI variants and related evaluation sets (CxN-NLI, CxN-NLI-Distinction), ensuring that improvements can be attributed to constructional abstraction rather than memorisation of specific lexical fillers.

Each construction is balanced across the three NLI labels, yielding 4,000 triples per construction and a total of 32,000 examples. The class balance ensures that macro-accuracy remains an unbiased measure of model performance.

4 Fine-Tuning Method

We use a small variety of base models for our fine-tuning experiments: Llama-3.1-8B-Instruct and Mistral-8B-Instruct. These models, both with approximately 8.1 billion parameters, have a decoder-only transformer architecture that has already undergone instruction tuning, making them proficient at following natural language prompts. Their relatively modest size allows for experimentation on single GPU setups, while their strong zero-shot baseline performance on tasks like NLI ensures that any observed gains from fine-tuning are both conservative and meaningful. We provide hyperparameters in Appendix C, full fine-tuning details in Appendix D, and training regimes in Appendix E.

5 Evaluation Framework

To rigorously assess our hypothesis that targeted fine-tuning yields systematic constructional understanding a comprehensive evaluation framework is employed. This framework specifies the core metrics for NLI tasks, outlines the use of diagnostic benchmarks to guard against overfitting and ensure generalization, and details essential controls and sanity checks to validate the genuineness of observed performance gains. Our primary metric to measure success is macro-accuracy. A statistically significant improvement of over 5% accuracy over a model's baseline evaluation (before fine-tuning) would be sufficient for us to accept our hypothesis.

5.1 Diagnostic Benchmarks

To ensure that improvements are not merely task-specific overfitting but represent genuine, transferable gains in understanding, performance on diagnostic benchmarks is critical. For this purpose we use Scivetti et al. (2025) the previously described CxN-NLI and CxN-NLI-Distinction datasets. These benchmarks feature out-of-distribution compositional tasks that involve the eight constructions targeted in fine-tuning; however, they are hand-crafted to test semantic understanding of the constructions.

If a model exhibits consistent performance across all of our datasets, yet remains consistent in these diagnostic benchmarks, we can confidently claim the model has improved on constructional usage; however, has not improved on the true understanding of the construction.

Model	Setting	CxN-NLI	CxN-NLI-Distinction
LLAMA-3.1-8B	baseline	57	33
	fine-tuned	66	39
Mistral-8B	baseline	49	36
	fine-tuned	63	37
GPT-4o	baseline	88	64
	3-shot ICL	91	65

Table 4: Results across CxN-NLI and CxN-NLI-Distinction benchmarks using baseline and ConTest-NLI fine-tuned models or in-context-learning (ICL) examples from ConTest-NLI.

6 Results and Discussion

ConTest-NLI demonstrated systematic gains across the CxN-NLI evaluation set; however, showed no improvements at semantic understanding of the CxN-NLI-Distinction dataset. Results are summarized in Table 4.

Ultimately, this shows model reasoning is done on surface-cues of constructions, rather than true constructional understanding. Critically, we know this is fundamentally different from human reasoning, where we are able to grasp the semantics of constructions instead of just surface-cues.

Notably, the Llama-3-8B-Instruct model, when fine-tuned on ConTest-NLI, showed a significant increase in accuracy on the CxN-NLI: from 57% to 66%. The Mistral-8B-Instruct model, with ConTest-NLI fine-tuning, saw performance rise from 49% to 63%. These improvements comfortably exceeded the hypothesized +5 percentage point threshold.

These ConTest-NLI results illustrate that fine-tuning on natural-sounding premises yields in-domain accuracy gains — key evidence that LLMs can internalize form-meaning mappings and structures when they encounter sufficiently varied, human-plausible NLI examples.

6.1 Error Analysis

We provide full error analysis in Appendix F. The six examples outlined in Table 8, each drawn from a different construction in the ConTest-NLI training set, illustrate the central weakness our paper identifies: the model's reliance on surface-level lexical and syntactic cues rather than robust, construction-general semantic reasoning. In each case, the model either (a) overfit to familiar lexical frames without integrating their semantic consequences, (b) failed to connect constructional form to the entailments it licenses (e.g., *way*-manner implying location change, resultatives implying caused state), or (c) ignored clear scalar or negation cues when they appeared in less frequent or slightly varied contexts. That these errors occur across all eight constructions — rather than being isolated to a single form — reinforces our quantitative finding: fine-tuning improved in-domain recognition but did not instill transferable, abstract constructional understanding.

6.2 Summary and Discussion

The fine-tuned model's improvement over the base model was 6% — above the 5% bar for the CxN-NLI evaluation set, but notably smaller than the gains on the CxN-NLI-Distinction dataset. This discrepancy implies that while the model does internalise certain abstract features of the constructions, a portion of the performance boost stems from adaptation to the repeated surface patterns and lexical distributions encountered during fine-tuning.

7 Future Work and Conclusions

Our investigation demonstrates that explicitly grounding LLM supervision in CxG yields measurable gains in systematic generalization, yet also exposes persistent limits of current models' abstraction capabilities. By fine-tuning small-scale LLMs on a CxG-informed corpus — ConTest-NLI — we show that targeted constructional supervision delivers substantial improvements (9% on existing CxG-NLI benchmarks), but that these gains attenuate on out-of-distribution and adversarial challenge items (CxN-NLI-Distinction dataset). These findings carry two broader implications for cognitive modeling.

First, our results suggest that, unlike human learners who extract and re-apply abstract form-meaning pairings across lexemes and structures, LLMs continue to rely on residual surface cues even after targeted fine-tuning. This divergence highlights the need for cognitive models of learning to account for both exemplar-driven acquisition and the development of schematic templates, offering a new benchmark against which to evaluate theories of human grammatical abstraction.

Second, the semi-automated pipeline we introduce — combining model-in-the-loop adversarial filtering and human validation — provides a scalable methodology for instilling constructional knowledge in models. Integrating such CxG-grounded datasets into training regimens can drive more robust semantic generalization, informing future architectures that more closely mirror human-like compositional reasoning.

References

Giulia ML Bencini and Adele E Goldberg. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of memory and language*, 43(4):640–651.

Claire Bonial and Harish Tayyar Madabushi. 2024. Constructing understanding: on the constructional information encoded in large language models. *Language Resources and Evaluation*, pages 1–40.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. OUP Oxford.

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

Adele E Goldberg, Devin Casenhiser, and Tiffani R White. 2007. Constructions as categories of language. *New ideas in psychology*, 25(2):70–86.

Adele E Goldberg and Ray Jackendoff. 2004. The english resultative as a family of constructions. *language*, 80(3):532–568.

Adele Eva Goldberg. 1992. *Argument structure constructions*. University of California, Berkeley.

Michael P Kaschak. 2007. Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition*, 35:925–937.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Steven Pinker. 2003. 16language as an adaptation to the cognitive niche. In *Language Evolution*. Oxford University Press.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025. Assessing language comprehension in large language models using construction grammar. *arXiv preprint arXiv:2501.04661*.

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

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*.

A Constructions of Focus

The constructions that we develop training data and test on are detailed in Table 5.

B ConTest-NLI Templates

Causative-With
- **Prompt:** Describe a situation where something causes a place or thing to have a new feature or quality.
  Example: The party filled the room with laughter and music.
- **Prompt:** Write about an action that makes an object filled or loaded with something else.
  Example: She packed the suitcase with clothes for the trip.
- **Prompt:** Imagine someone causing a change by adding something to a space. Describe it.
  Example: They stocked the pantry with canned goods before the storm.

Caused-Motion
- **Prompt:** Write about someone making something move to a new place.
  Example: He pushed the broken car into the garage.
- **Prompt:** Describe an action that results in an object being relocated somewhere.
  Example: She threw the ball across the yard.
- **Prompt:** Tell a short story where an action causes an object to end up somewhere else.
  Example: The wind carried the leaves onto the porch.

Comparative-Correlative
- **Prompt:** Describe a situation where two things change together — as one increases or decreases, so does the other.
  Example: The more he practiced, the better he became at playing the piano.
- **Prompt:** Write a sentence showing how more or less of one thing affects another thing.
  Example: The less she slept, the grumpier she got.
- **Prompt:** Imagine a cause-and-effect relationship where two actions or qualities are linked. Explain it.
  Example: The more it rained, the faster the river rose.

Conative
- **Prompt:** Write about someone trying to interact with an object but not necessarily succeeding fully.
  Example: He tugged at the door, but it wouldn't budge.
- **Prompt:** Describe an action where a person touches or tries to affect something without completely changing it.
  Example: She tapped at the microphone to check if it was working.
- **Prompt:** Imagine someone fiddling with or attempting to do something to an object — describe it.
  Example: He poked at the firewood, trying to get the flames to grow.

Intransitive Motion
- **Prompt:** Describe a person, animal, or thing moving from one place to another.
  Example: The cat wandered into the kitchen.
- **Prompt:** Write about a movement where the focus is on someone or something changing location.
  Example: The children raced down the hill.
- **Prompt:** Tell a short story about a journey or movement from one place to another.
  Example: The balloon drifted across the blue sky.

Let-Alone
- **Prompt:** Describe a situation where something is already hard or unlikely — and an even harder thing is even

Construction	Example Sentence
LA (Let Alone)	I can't knit a scarf, *let alone sew a quilt*.
CC (Comparative Correlative)	The faster you run, the sooner you tire.
CWT (Caused Motion with Theme)	She filled the bucket with sand.
CON (Conative)	The boxer *punched at* the heavy bag.
WAY (Way Construction)	She danced her way to fame.
IM (Intransitive Motion)	The children ran into the park.
CM (Caused Motion)	They rolled the log into the river.
RES (Resultative)	He hammered the metal flat.

Table 5: Eight challenge constructions ordered from most lexically substantive (top) to most schematic (bottom). Each example instantiates the construction in context.

less likely.
Example: He couldn't finish a page of his homework, let alone the entire assignment.

- **Prompt:** Write about two related actions or qualities, where the second is even more extreme than the first.
Example: I can barely manage to jog a mile, let alone run a marathon.
- **Prompt:** Imagine someone struggling with one task — and an even harder task is even more impossible. Describe it.
Example: She had trouble cooking pasta, let alone baking a soufflé.

Resultative

- **Prompt:** Describe an action that causes something to change its state or condition.
Example: He wiped the counter clean.
- **Prompt:** Write about an event where someone does something that makes an object end up different than before.
Example: She hammered the metal flat.
- **Prompt:** Tell a story where an object transforms because of someone's actions.
Example: They painted the walls bright yellow.

Way-Manner

- **Prompt:** Describe someone making progress by doing an action repeatedly or in a special way.
Example: He elbowed his way through the crowded hallway.
- **Prompt:** Write about someone moving through space by performing an activity along the way.
Example: She laughed her way down the mountain trail.
- **Prompt:** Imagine someone reaching a destination while doing something unusual — describe it.
Example: They danced their way to the front of the stage.

C Hyperparameters

Hyperparameters and justifications are given in Table 6.

D Fine-Tuning Details

Given the size of our labeled fine-tuning data, full fine-tuning of all model parameters would be computationally expensive, prone to overfitting, and very inflexible for our experiments. Therefore, we employ LoRA; This approach significantly mitigates the risk of overfitting on smaller, highly structured datasets like ours.

LoRA modules (rank r=16, scaling factor α=32, dropout p=0.05) are specifically injected into the attention layers and multi-layer perceptron projections of layers 12 through 20 of the Llama-3-8B-Instruct model.

Layers 12-20 in a 32-layer transformer model, such as Llama 3.1 8B, are roughly in the middle of the network. Prior research shows that middle and upper-middle layers often encode a mix of syntactic and semantic abstractions - ideal for adapting models to semantic tasks like NLI, especially for constructional generalization (Liu et al., 2019).

We also note that training is conducted using mixed-precision, with weights in BFLOAT16 and

Parameter	Value	Justification
`temperature`	0.8	maximises lexical variety without destabilising syntax
max tokens	500	covers premise + three hypotheses with margin
rare-lemma seed list	5 376 nouns/verbs/adjectives	reduces overlap with pre-training corpora

Table 6: Generation parameters and their justification.

activations in INT8, to further reduce memory footprint and improve training efficiency. Additional fine tuning hyperparameters are found in Table 7.

E Training Regimes

To systematically investigate how and where constructional knowledge is acquired and represented, three distinct training regimes are employed. These regimes are designed to disentangle the effects of weight updates, classifier head architecture, and dataset characteristics.

Shared-Head: Updates the model with a single shared three-way NLI head for all constructions. This is the canonical regime, testing if a unified representation can be learned across all constructions.

Per-construction Heads: Updates the model with 8 independent NLI heads (one per construction). This explores whether separate, specialized classifier heads better capture constructional nuances.

In-Context Few-Shot: No weights are updated. Predictions are made via prompting (8-shot). This baseline tests learning from examples in context, without training.

All fine-tuning regimes are run for a maximum of 5 epochs over the training data. By comparing performance across these regimes, we can draw more nuanced conclusions: for example, if the Shared-Head regime significantly outperforms In-Context Few-Shot, it suggests that explicit weight updates are beneficial. The Per-construction Heads condition offers insights into the potential modularity of learned constructional knowledge. This comprehensive experimental design ensures that claims about improved constructional understanding are robust and well-substantiated.

F Full Error Analysis

The examples extracted and displayed in table 8 each illustrate a distinct type of model failure identified in our study.

In the Conative and Way-manner cases, the model recognised the action but failed to apply constructional entailments — ongoing effort should contradict "gave up", and the Way construction implies location change. The Caused-motion and Resultative examples show that the model often conflates transformation events with generic processes, ignoring the causative semantics that the construction encodes. The Let-alone example reveals a missed scalar inference, treating "barely managed" as isolated from the second clause. Finally, the Intransitive-motion case highlights a negation cue failure, where "without a destination" was incorrectly aligned with a positive statement due to lexical overlap. Across constructions, these failures demonstrate that improvements from fine-tuning largely reflect memorisation of surface patterns rather than abstraction of form–meaning pairings.

Optimizer and Hyperparameters	
Optimizer	AdamW
β_1	0.9
β_2	0.999
ϵ	$1e-8$
Regularization and Early Stopping	
BookCorpus in minibatches	33% (for anti-forgetting)
BookCorpus loss weight	0.25
Label smoothing (NLI heads)	0.1
Gradient clipping (max norm)	1.0
Weight decay (LoRA matrices)	0.01

Table 7: Hyperparameters from fine-tuning experiments

Construction	Premise	Hypothesis	Gold Label	Model Label
Conative	The carpenter repeatedly hammered at the stubborn nail.	The carpenter gave up trying to fix the nail.	Contradiction	Neutral
Way-manner	The detective elbowed his way to the front of the crowded room.	The detective stayed at the back of the room.	Contradiction	Neutral
Caused-motion	The artist painted the mural into a vibrant masterpiece.	The artist worked on the mural for a week.	Neutral	Entailment
Let-alone	He barely managed to tie his shoelaces, let alone complete the marathon.	He found tying his shoelaces easy.	Contradiction	Neutral
Intransitive-motion	Without a destination, the traveler wandered through the forest.	The traveler had a clear destination in mind.	Contradiction	Entailment
Resultative	A few strikes were enough: the blacksmith hammered the iron flat.	The iron became flat.	Entailment	Neutral

Table 8: Examples of failed NLI cases from the ConTest-NLI training set.

Construction Grammar Evidence for How LLMs Use Context-Directed Extrapolation to Solve Tasks

Harish Tayyar Madabushi
University of Bath, U.K.
htm43@bath.ac.uk

Claire Bonial
DEVCOM Army Research Laboratory, U.S.A.
claire.n.bonial.civ@army.mil

Abstract

In this paper, we apply the lens of Construction Grammar to provide linguistically-grounded evidence for the recently introduced view of LLMs that moves beyond the 'stochastic parrot' and 'emergent Artificial General Intelligence' extremes. We provide further evidence, this time rooted in linguistic theory, that the capabilities of LLMs are best explained by a process of *context-directed extrapolation from their training priors.* This mechanism, guided by in-context examples in base models or the prompt in instruction-tuned models, clarifies how LLM performance can exceed stochastic parroting without achieving the scalable, general-purpose reasoning seen in humans. *Construction Grammar is uniquely suited to this investigation, as it provides a precise framework for testing the boundary between true generalization and sophisticated pattern-matching on novel linguistic tasks.* The ramifications of this framework explaining LLM performance are three-fold: first, there is explanatory power providing insights into seemingly idiosyncratic LLM weaknesses and strengths; second, there are empowering methods for LLM users to improve performance of smaller models in post-training; third, there is a need to shift LLM evaluation paradigms so that LLMs are assessed relative to the prevalence of relevant priors in training data, and Construction Grammar provides a framework to create such evaluation data.

1 Introduction

Understanding how Large Language Models (LLMs) solve complex tasks is a critical yet unsettled question, and the field remains divided between two primary viewpoints. One perspective characterizes LLMs as 'stochastic parrots,' which do little more than generate statistically probable outputs based on their training (Bender et al., 2021; Bender and Koller, 2020; Mitchell and Krakauer, 2023). The opposing view contends that with sufficient scale in parameters and data, LLMs exhibit 'emergent reasoning' (Brown et al., 2020a; Wei et al., 2022b; Srivastava et al., 2023a), a phenomenon claimed to be 'sparks of Artificial General Intelligence' (AGI) (Bubeck et al., 2023).

Our recent work (Tayyar Madabushi et al., 2025b) has sought to bridge this divide with an alternative framework.[1] Rather than viewing LLMs as either 'stochastic parrots' or as possessing advanced, human-like reasoning, we contend that the capabilities and limitations of these models are best explained by **context-directed extrapolation from their training priors.** In our framework, the necessary context is supplied by in-context learning examples for base models, or directly by the prompt for instruction-tuned models.

This position paper first summarizes the framework proposed in Tayyar Madabushi et al. (2025b) (Section 2). We then present our working definition of reasoning and generalization while providing linguistic examples of the generalization of constructions (Section 3). We discuss the two prevalent views of LLM capabilities along with evidence from CxG rsearch for and against each view. First, we explore stochastic parroting and present evidence of LLM success in solving difficult, non-memorizable problems that require more than next-token prediction (Section 4). Second, we explore the possibility of AGI, where we present research demonstrating that models are incapable of completing certain tasks that are trivial for humans (Section 5). This pattern, we will argue, suggests a specific shortcoming in what is termed 'advanced reasoning.' We then present new evidence from Construction Grammar (CxG) that substantiates this view (Section 6) and provides insights into the limitations of the more extreme, alternative views.

[1] Mentions of our past research have been de-anonymized after double-blind review and paper acceptance.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 190–201

From this foundation, we turn to the problem of evaluation. We argue that even though LLMs have mastered many superficial linguistic elements, sound linguistic theory provides the necessary tools to test their deeper reasoning (Sections 7, 8). *Specifically, we demonstrate how the principles of CxG can be used to design precise tests that probe the inherent capabilities of these models, and we suggest extensions informed by usage-based theories.*

2 Context-Directed Extrapolation from Training Priors

In the framework of context-directed extrapolation, an LLM makes use of the entire prompt context to generate its output. This process is straightforward in base models, which are trained exclusively on the next-token prediction objective. For base models, the input prompt provides the sequence context from which the most probable subsequent token is generated. However, dealing with the more common models, which are additionally trained to follow instructions (instruction-tuned models), the instructions in the prompt establish a *semantic context*. This context is then used to extrapolate from relevant priors acquired during pre-training, as opposed to treating the prompt merely as a token sequence.

Specifically, for base models, while there is wide debate over how LLMs function, their capabilities and their ability to truly generalize, their capacity for *in-context learning (ICL)* is an indisputable fact (Brown et al., 2020b; Olsson et al., 2022). ICL is an ability of LLMs to learn a new task on the fly, simply by being given a few examples within the prompt. To illustrate this, Tayyar Madabushi et al. (2025b) use the example of a modified addition task. In this task, when provided with the input prompt:

$$1 + 3 = 5; 7 + 12 = 20; 8 + 3 =$$

LLMs, trained only on the next token prediction task, can infer the novel pattern $(a + b + 1)$ from the examples and produce the correct, non-obvious answer of 12.

In Tayyar Madabushi et al. (2025b), we derive the notion that ICL is a method of solving tasks by extrapolating from pre-training priors from a convergence of several distinct theories. We note that research consistently supports this view, whether by directly linking ICL to the distributions in pre-training data (Chan et al., 2022; Hahn and Goyal, 2023), or by explaining it through frameworks like Bayesian inference (Zhang et al., 2023; Xie et al., 2021) and Probably Approximately Correct (PAC) learning (Li et al., 2023b). This conclusion is reinforced by other studies that liken ICL to fine-tuning (Dai et al., 2023) or show that it can implicitly perform gradient descent, a process linked to meta-learning (Akyürek et al., 2023; Li et al., 2023a; Zhang et al., 2024; Von Oswald et al., 2023). Ultimately, we argue that regardless of the specific mechanism, all existing research indicates that ICL fundamentally relies on priors from pre-training data, with the in-context examples serving to guide the model toward the relevant priors needed for the task at hand.

2.1 Context-Directed Extrapolation in Base vs Instruction-Tuned Models

A critical observation is that LLMs trained solely on next-token prediction (i.e. base models) are by construction nothing more than sequence completion engines. However, *these base models cannot solve tasks that require abstract reasoning without being provided with examples through in-context learning (ICL)* (Lu et al., 2023). Consider, for instance, the following logical deduction problem from the Big-Bench benchmark:

> Question: On a shelf, there are five books: a gray book, a red book, a purple book, a blue book, and a black book. The red book is to the right of the gray book. The black book is to the left of the blue book. The blue book is to the left of the gray book. The purple book is the second from the right.
>
> Targets: 'The gray book is the leftmost.': 0; 'The red book is the leftmost.': 0; 'The purple book is the leftmost.': 0; 'The blue book is the leftmost.': 0; 'The black book is the leftmost.': 1

Base models fail on such reasoning tasks when presented without examples, however, they can solve this task when presented with a prompt that includes examples. Central to our augment is the fact that, instruction-tuned models can solve this task without examples based purely on a description of the task (Lu et al., 2023).

Context-directed extrapolation from training data priors offers a unifying framework to explain both the capabilities and, importantly, the limitations of LLMs: In base models, the in-context

examples provide the context direction to allow the model to infer and solve the relevant task at hand. In instruction-tuned models, however, the process of instruction tuning allows the models to interpret the semantic context of the prompt without explicit examples, and similarly direct extrapolation. We contrast our framework with that of stochastic parroting in Section 4.1.

2.2 Extrapolation and Grounding

An important implication of context-directed extrapolation is that *it allows for a limited form of grounding.* By this we do not mean that models achieve grounding in the human sense of connecting language to embodied experience. Rather, because the mechanism involves extrapolating from priors activated by the prompt, information that is not explicitly present in surface form can nevertheless become available to the model. For example, when confronted with a nonce verb whose definition is provided in the prompt, the model can project that meaning into novel contexts and apply it productively. Indeed, this same process allows models to respond effectively in tasks such as the Sally–Anne test (Wimmer and Perner, 1983), enabling models to succeed on certain Theory of Mind evaluations that would be inaccessible to 'stochastic parroting' (Kosinski, 2024).

This is categorically different from stochastic parroting. A purely parroting mechanism cannot accommodate genuinely novel input that falls outside its memorized distribution. The fact that LLMs can extend prompt-based definitions, apply abstract patterns, and generate context-appropriate interpretations indicates that extrapolation yields access to extrapolatable information that is not reducible to surface statistics. In this sense, context-directed extrapolation provides a pathway to limited grounding, albeit one constrained by the priors in training data and the context supplied at inference time.

2.3 A Mechanistic Basis for Context-Directed Extrapolation

To understand the underlying mechanics of this capability in LLMs, in Tayyar Madabushi et al. (2025b), we first point to the foundational work of Olsson et al. (2022), who systematically showed that LLMs could complete abstract patterns with random tokens (e.g., given a sequence $[A][B]...[A]$, LLMs correctly respond with $[B]$). While this compellingly refutes the 'stochastic parrot' notion by suggesting an algorithmic capability, we introduce a crucial caveat from recent research (Niu et al., 2025): this pattern-matching ability degrades significantly as the tokens become less frequent in the pre-training data. This finding demonstrates that even this seemingly abstract skill is fundamentally tethered to the model's training priors.

We then argue that this powerful, data-dependent pattern-matching ability is the same core mechanism that allows LLMs to solve more complex tasks via ICL. This view is substantiated by evidence showing that ICL remains effective even when the labels in the examples are manipulated, such as being flipped between positive and negative or replaced with entirely unrelated words like 'Foo' and 'Bar' for a sentiment classification task (Wei et al., 2023). Therefore, in Tayyar Madabushi et al. (2025b), we conclude that ICL, while impressive, is a sophisticated but ultimately constrained process. We argue that because its operation is always guided by the user-provided examples and bound by the limits of its training data, it fails to meet the requirements for advanced, generalizable reasoning. In this setting, the model never gains true 'agency,' as its performance is always a function of the input, preventing it from making the leap from guided pattern-matching to unguided, human-like cognition.

3 Construction Grammars and Generalization

In this section, we outline our definition of human-like reasoning and provide insights into such reasoning in linguistic settings from CxG. Following our work in Tayyar Madabushi et al. (2025b), we embrace a definition of advanced reasoning that requires mastery and understanding of knowledge taken from one set of members instantiating a class, and then generalization and application of that knowledge to a novel set of items. In terms of CxG, constructions (defined as pairings of meaning and form at any level—morphological, lexical, phrasal (Goldberg, 2003; Hoffmann and Trousdale, 2013)) should be thought of as classes, and members are certain instantiations or realizations of that construction/class. Psycholinguistic evidence from child language acquisition demonstrates that children acquire frequently-heard constructions first and initially only use the member instantiation that they have heard (Tomasello, 2009). For example, a child's first Resultative construction will likely involve the high-frequency verb "make" along with

other lexical items the child is frequently exposed to: "Mommy made me mad." An 'understanding' of this construction is achieved when a speaker can recognize the similarity of other instantiations of this construction, which generally involve some kind of verb of change-of-state semantics within the structure (e.g., "Berries turned me blue!"). True generalization of the construction requires abstracting and applying knowledge of the construction from heard instantiations to novel items—in this example, novel instantiations of the phrasal constructions where the individual lexical items have likely not been experienced within that construction before: e.g., "The dog barked me awake."

Over the next sections, we will discuss the more extreme viewpoints of LLM performance as either "stochastic parrots" or advanced general intelligence. In each section, we will close with relevant research from CxG. Our review of work on CxG will reveal a mixed picture: models can make the required generalization in some instances, but fail in others. However, based on our framework of context-directed extrapolation, these seemingly contradictory performances become explainable.

4 LLMs are NOT Stochastic Parroting

While the 'Stochastic Parrots' paper from Bender et al. (2021) rightly identifies the risks of bias propagation in large-scale models, its claim that these models merely generate the next most likely token is demonstrably false, as we will show. We define stochastic parroting as the mechanism of generating the *precise* statistically most likely next token given the immediate input sequence. In this view, an instruction is merely more text to be completed. In the following sections we contrast this view, with our view that LLMs solve tasks using context-directed extrapolation from training priors.

4.1 Stochastic Parroting vs. Context-Directed Extrapolation

Functional Commonalities. *From the perspective of the performance of base models, there is no functional difference between context-directed extrapolation and stochastic parroting.* Base models consistently fail tasks such as the one described in Section 2.1 when presented without examples. One can argue that the examples simply form a long context, where the correct answer is the most probable sequence completion. This makes both theories appear to describe the same mechanism: the model completes a given sequence based on statistical patterns. Consequently, the two views are indistinguishable when analyzing this alone.

Most LLMs in wide use, such as public chat models, undergo instruction fine-tuning after their initial pre-training so they can 'understand' and follow instructions presented within their prompts (Wei et al., 2022a). This additional training, however, complicates their evaluation. It becomes difficult to tell whether a model's success on a new task is a sign of genuine emergent reasoning or simply a consequence of its training on similar tasks.

This issue was explored in a systematic study by Bigoulaeva et al. (2025), who fine-tuned over 90 models and demonstrated that the performance of instruction-tuned models is strictly correlated with that of base models. This suggests a single underlying mechanism is at play in both. Building on this, in Tayyar Madabushi et al. (2025b), we argue that this mechanism is context-directed extrapolation from pre-training data. We propose that *instruction-tuning simply allows the model to perform the same kind of extrapolation from a natural language prompt, rather than needing the explicit in-context examples that base models require.*

Functional Difference. The functional difference between these two views becomes apparent with instruction-tuned models. A base model, provided with enough examples, generates the correct output because it becomes the most probable completion of that long sequence. In contrast, the context-directed extrapolation view posits that instruction-tuning enables a different mechanism. It allows the model to interpret an instruction not as a literal sequence to be continued, but as a directive to construct an implicit context for a task. This allows the model to activate relevant priors (just as examples do for base models) from its pre-training data to perform the task specified by the prompt, rather than simply completing the text of the prompt itself. *Critically, the evidence for this distinction is that instruction-tuned models can solve the logical deduction (and similar) problems presented in Section 2.1 without any examples (Lu et al., 2023). This phenomenon cannot be explained by stochastic parroting, but is directly accounted for by context-directed extrapolation.*

This distinction becomes even more stark in tasks involving novel words, as this eliminates the model's ability to rely on pre-existing statistical associations. The Winodict benchmark (Eisensch-

los et al., 2023), for instance, modifies Winograd schemas by replacing a critical verb with a nonce word defined within the prompt. Consider:

> The verb 'to plest' means to be scared of... The city councilmen refused the demonstrators a permit because they **plested** violence."

To correctly resolve the pronoun "they," the model cannot use any stored knowledge about the word "plest." It must parse the definition provided in the prompt and apply that meaning to the sentence. The success of models on this task provides compelling evidence that the model is not merely predicting a statistically likely token, but is using the in-prompt definition to build a context and reason accordingly. The ability of LLMs to successfully solve this task is directly explained by context-directed extrapolation as it allows models to extropolate meaning from context. In contrast, a pure stochastic parroting mechanism based on predicting the next likely token along *cannot* account for this ability. As discussed previously, unlike base models (Section 2), instruction-tuned models succeed on tasks such as logical deduction without explicit examples (Section 4.1), a result that cannot be explained by stochastic parroting. The Winodict benchmark illustrates this distinction especially clearly. By replacing a key verb with a nonce word defined only within the prompt, the task prevents the model from relying on stored associations. Yet models are still able to resolve the pronoun correctly by projecting the definition into novel contexts (Section 6), a behavior that cannot be accounted for by a purely stochastic parroting mechanism. Indeed, mechanistic studies exploring 'induction heads' further support this view (Section 2.3). In what follows, we turn to CxG research relating to the notion of stochastic parroting.

4.2 CxG & Stochastic Parroting

There is relevant research demonstrating first that information on certain constructions is present in pre-training data, such that models may rely on stochastic parroting to provide the impression of proficiency with the constructions of the language. Tayyar Madabushi et al. (2020) probe a variety of BERT-based models for access to knowledge of several constructions proposed in Dunn (2017). In this work, Tayyar Madabushi et al. (2020) test BERT models on their ability to distinguish sentences that are instances of a given construction from those that are not. Alongside the base model, the authors trained several BERT "clones" with additional exposure to constructional information, varying the frequency of constructions during pre-training so that some clones saw high-frequency items and others saw low-frequency ones. The expectation was that clones trained on rarer constructions would benefit most, since such items were unlikely to appear often in the original pre-training data. However, the results showed little improvement over the base BERT model, leading the authors to conclude that constructional knowledge was already accessible to BERT. It is worth noting, though, that the constructions targeted were identified in a data-driven way using the methods proposed by Dunn (2017), and typically involved fixed lexical items. More schematic phrasal patterns, such as argument structure constructions (Goldberg, 1992), were not included. As a result, it is plausible that the constructions tested were already present in the base model's pre-training corpus at sufficient frequencies to allow strong performance through context-directed extrapolation rather than deeper generalization.

In Bonial et al. (2025), we provide evidence from Multi-Word Expression (MWE) constructional templates, which demonstrate that LLM abilities must go beyond stochastic parroting. While it is expected that the conventional meanings of common MWEs such as "spill the beans" (meaning *to tell a secret*) would be encoded in LLMs, we show that they can use in-context learning and a single example to teach models entirely novel MWEs, such as "winking at pringles" (meaning *to indulge in frivolity*). Not only can the models explain the meaning of a new (non-verbatim) usage of the novel MWE, GPT-4o is able to answer targeted questions about the new usage, for example that "a meeting where everyone is winking at a bunch of chips" is probably not a productive use of time. Thus, this research demonstrates that ICL in the form of a definition of the novel MWE and single usage example facilitates performance that cannot be dismissed as stochastic parroting, as these novel MWEs are guaranteed not to be found in pre-training data. Table 1 presents further examples from this work.

5 LLMs are also NOT AGI

On the other end of the spectrum, some research argues that models, when scaled up, gain abilities

MWE Definition & Example	Novel Usage	GPT4o Interpretation	Annotator Interpretation
Expression: **to send one's LLM (instead)** Meaning: Not worth one's time. Usage: There's no way I'm attending that meeting, I'll probably send my LLM	The bank teller seems to have sent his LLM to work today.	The statement means that the bank teller seems disengaged or inattentive, implying they are not putting in effort or paying attention, as if they had metaphorically 'sent their LLM' to do their job instead of being present themselves.	Annotator 1: The bank teller is not putting effort into their work, relying on rote memorization of their tasks, thereby making many mistakes and potentially misunderstanding some of their interactions.
Expression: **dance with the deer** Meaning: to have a deep and meaningful conversation, especially about something one is passionate about. Usage: When they were planning their research, they often danced with the deer.	I don't think they were dancing with the same herd of deer, possibly not even the same species.	The statement implies that the individuals involved were having deep and meaningful conversations, but their passions, ideas, or perspectives were significantly different—perhaps even fundamentally incompatible. It suggests a lack of alignment in their discussions or goals.	Annotator 1: Although all parties are interested and passionate about their topic of conversation, they have not reached common ground as to what the underlying meaning and implications of what the others are saying.

Table 1: Sample evaluation results from Bonial et al. (2025), in which we tested the ability of LLMs to generalize to novel MWEs, given and defined in the first column. Note that models could not have been exposed to these specific MWEs during pretraining, yet the interpretation of the novel usage (second column) is quite similar to that of human annotators.

akin to high-level human reasoning (Brown et al., 2020b; Wei et al., 2022b; Srivastava et al., 2023b; Lu et al., 2024; Wei et al., 2024).

In Tayyar Madabushi et al. (2025b), we argue that high-level reasoning is demonstrated only when a model solves tasks it was not explicitly trained for, distinguishing genuine cognitive application from simpler forms of understanding (Krathwohl, 2002). In line with Chollet (2019), we note that a model trained solely to master a single task such as chess, even to a superhuman level, does not exhibit the kind of reasoning that matters here, since it is not generalizing knowledge to a truly new domain. To make this distinction precise, here and in Tayyar Madabushi et al. (2025b), we adopt the framework of Krathwohl (2002), a revision of Bloom's original taxonomy of educational objectives (Bloom et al., 1956), which defines advanced reasoning as the ability to apply and extend knowledge beyond familiar instances to novel contexts.

To argue that LLMs are not performing advanced reasoning, we point to two key shortcomings: models' tendency for hallucination and their failure on seemingly simple tasks. First, LLM hallucinations—outputs that are not aligned with reality—are cited as a major piece of evidence against advanced reasoning (Huang et al., 2025). We argue this phenomenon should not be confused with human confabulation, as there is no evidence for LLM agency (Lu et al., 2024), and these errors can be traced to the model defaulting to statistical patterns from its training data when the prompt's context is insufficient (Hanneke et al., 2018). Second, we highlight that LLMs often fail at tasks that are trivial for humans (Nezhurina et al., 2025). For instance, even top models perform poorly on clinical psychology *faux-pas* tests compared to children (Shapira et al., 2023), and they are significantly outperformed by non-expert humans in simple AI planning domains like Blocksworld (Valmeekam et al., 2023).

5.1 CxG & Advanced Reasoning

From a constructional perspective, 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. In their design, the authors adopt a sorting task where both human participants and models are asked to judge sentence similarity. The dataset is deliberately constructed so that the constructions under investigation are expressed through a range of lexical verbs. Importantly, the verbs chosen to instantiate different constructions belong to overlapping semantic classes—for instance, verbs such as *cut* and *slice*. This setup allows them to test if participants and models cluster sentences on the basis of verb meaning, as traditional generative grammar would suggest, or if they recognize the broader constructional pattern. The findings reveal a sharp divergence depending on model scale. MiniBERTas

(Warstadt et al., 2020), a model with only one million parameters, aligns sentences primarily by verb-level semantics, whereas the much larger RoBERTa model (30B parameters; (Liu et al., 2021)) instead groups them in line with constructional semantics. While the authors do not point to this as evidence of advanced reasoning *per se*, they do conclude that larger models perform like native speakers while smaller models perform more like second language learners. However, we emphasize that these results can also be interpreted as larger models successfully extrapolating from pre-training priors that the smaller models do not have.

Additional studies using CxG highlight similar limits to the reasoning abilities of LLMs. Weissweiler et al. (2022a) examine the Comparative-correlative construction (e.g., The higher you fly, the harder you fall) as a test case for whether models can capture both its syntactic properties and its associated semantic meaning. Their methodology first targets the syntax by evaluating whether models can reliably recognize instances of the construction in natural corpus data and in controlled, synthetic examples. On this task, several BERT-based models perform well, successfully identifying and discriminating the construction. Such results are not unexpected given that the Comparative-correlative includes fixed lexical items in key structural positions. The crucial question, however, is whether models can also handle the semantics of the construction. To probe this, the authors evaluate performance on a downstream task that requires reasoning about the correlational meaning encoded by the construction. Here the models perform poorly, especially on nonce words, with accuracy barely above chance, indicating that while BERT-based models can recognize the formal template of the Comparative-correlative, they fail to grasp its interpretive content. We highlight that this failure on nonce words is, yet again, indicative of context-directed extrapolation. Similar research evaluating both formal recognition and semantic interpretation of the Causal-excess construction underscores this finding—models can pick out the construction but perform poorly on semantic understanding tests in the form of downstream questions (Zhou et al., 2024).

6 CxG & Context-Directed Extrapolation

In Bonial and Tayyar Madabushi (2024a), we find that even the largest models available at the time (GPT-3.5 and GPT-4) are restricted to recognizing substantive constructions (with fixed words), whereas schematic constructions (without fixed words) elude recognition of either form or meaning. In that research, we collect and leverage the CoGS dataset (Bonial and Tayyar Madabushi, 2024b), which includes approximately 500 corpus instances of 10 different phrasal constructions of varying schematicity (i.e. some constructions are fully fixed words, while others are argument structure constructions with no fixed words). The corpus includes relatively frequent constructions, but is limited to instantiations of those constructions that are not the most frequent, entrenched instantiations. For example, the Ditransitive construction instances do not include usages with the verb "give," which is the most frequent verb to instantiate this construction: "He gave me a book." Instead, CoGS Ditransitives include only cases where the lexical semantics of the instantiating verb do not inherently include transfer semantics: "He poured her a martini." In other words, the constructions in CoGS have high type frequency, but these particular instantiations have relatively low token frequency. Nonetheless, the fixed words of the substantive constructions facilitate tapping into the appropriate pre-training data in order to recognize the construction (but not necessarily a deeper understanding, as suggested by (Weissweiler et al., 2022b)). In contrast, although the schematic argument structure constructions are the most fundamental constructions of the English language with very high type frequency (Goldberg, 1992), the models are not able to apply generalized formal and semantic properties of the construction to novel instantiations. This suggests that models can extrapolate to a point to account for relatively infrequent, creative instantiations of constructions, but the level of generalization required for recognizing the structural slots and associated semantics of argument structure constructions is beyond model abilities.

Similarly, Scivetti et al. (2025a) find that the "human-scale" BabyLM demonstrates strong formal knowledge of the Let-alone construction, but no understanding of the associated scalar semantics. Further experiments on the templated evaluation dataset first remove all Let-alone constructions from pre-training data, as well as filtering all related constructions (e.g., Much-less). The authors find that this does not change BabyLM performance on formal recognition of the construction.

The authors then remove all individual "let" and "alone" tokens from pre-training, and this significantly degrades performance on formal recognition, leading us to conclude that the model is drawing on compositional, lexical information of the individual words as opposed to the form of the phrasal construction as a whole. Thus, this research underscores the notion that generalizing the semantics associated with syntactic slots of a construction eludes models, and casts further doubt on whether or not even the formal features learned by models are generalized at the constructional level or limited to lexical, collocational features.

In Scivetti et al. (2025b), we provide further evidence of both the extrapolation abilities of LLMs when it comes to constructions, as well as the limits of their generalization abilities. We leverage a subset of the CoGS dataset described previously, specifically using corpus constructional usages as the premises for Natural Language Inference (NLI) triples in which templates are leveraged to semi-automatically generate entailed, neutral, and contradicted hypotheses. In leveraging an NLI task, we test downstream, functional understanding of the CoGS constructions, which again are of relatively high type frequency (e.g., Ditransitive: "he gave me a book") but the instantiations of those constructions are relatively low token frequency (e.g., Ditransitive: "he poured her a martini"). Interestingly, although we found that models failed to recognize more schematic constructions in Bonial and Tayyar Madabushi (2024a), in our subsequent NLI research (Scivetti et al., 2025b), we find that the largest models available (GPT-4 and 4o) perform comparably on the constructional NLI and Stanford NLI, ostensibly demonstrating that the models are able to draw inferences correctly over the constructional premises.

In Scivetti et al. (2025b), we then conduct follow-on experiments where the models are evaluated on a new set of NLI triples involving schematic constructions that are not the high type-frequency constructions of CoGS but are formally indistinguishable. For example, the Depictive construction (e.g., "She bought the apples fresh") has the same syntactic slots as the Resultative (e.g., "She hammered the metal flat"), but distinct semantic roles associated with the slots. We then test the same models on NLI triples involving the formally identical but semantically distinct premises, and find that model performance drops substantially. We posit that this research therefore demonstrates the limits of extrapolation as opposed to true generalization of the meaning of constructions. The strong performance on the original CoGS premises shows that models can effectively extrapolate from pre-training data, which is ample for these high type-frequency constructions. However, the degradation in performance on the formally identical but semantically distinct premises shows that because models are extrapolating from the higher-frequency constructions, they will perform the task (incorrectly) according to those priors when faced with lower-frequency constructions that the model seems unable to distinguish.

Finally, in the second set of results from Bonial et al. (2025), we extend this line of evidence. We show that while LLMs can learn and use entirely novel MWEs when definitions are provided in the prompt (as discussed in Section 4.2, see also Table 1), performance degrades when models are asked to reason across multiple MWEs at once. For example, given novel MWE definitions for "drown the cables" (an invented MWE defined as *to sever or overwhelm communication*) and "dance with the deer" (*to have a deep, meaningful conversation*), the models were evaluated for their ability to reason about the semantic interaction of the two MWEs in a novel usage involving both MWEs. Human annotators were able to do this consistently, but even advanced models like GPT-o1 and GPT-4o faltered. This demonstrates the limits of context-directed extrapolation, which enables models to extend clear, explicit definitions to new usages (as shown in this work for single MWE), but that the mechanism struggles once the links between constructions become less direct.

7 Discussion and Implications

Context-directed extrapolation explains LLM behavior as the use of priors activated by prompt context. Because of this, the very same capability, such as apparent Theory of Mind, will be observed when the relevant priors are strong, but absent or much weaker when priors are sparse. The same holds for grounding: it will appear when relevant information is easily extrapolatable from context and fail when it is not. This means that evaluation must carefully distinguish between cases where models are simply drawing on rich priors and cases where success would require true human-like generalization. Counterfactuals are ideal for making this

distinction, since they force the model to reason beyond memorized or extropable priors, and LLMs consistently fail on such tests despite succeeding on superficially similar ones (Wu et al., 2024).

For decades NLP research sought to build pipelines around symbolic templates and formal reasoning systems. Over time the pipeline itself became an end goal. LLMs now shift this landscape by allowing us to fill templates more easily and then use established resources, such as AMR (Banarescu et al., 2013; Bonial et al., 2018) or frame semantics (Fillmore et al., 2012), to support reasoning processes in systematic, verifiable ways (e.g., Tayyar Madabushi et al. (2025a)). *Given that models continue to struggle with more advanced reasoning tasks, it is increasingly important to see them as an interface between the complexity of language and downstream formal reasoning rather than as reasoning systems themselves.*

CxG is a particularly strong testbed for this view. It allows us to probe the line between semantics and syntax and to see where models succeed because of exposure to canonical patterns of language and where they fail to generalize. Because there is already extensive evidence of how humans learn and extend constructions (e.g., Tomasello (2009)), CxG provides the right framework to compare human generalization against model extrapolation and to identify the precise gaps that remain. Usage-based theories of learning, such as Frame Semantics (Fillmore et al., 2012), can also be incorporated into the design of systems. We need an interface between the lexical, surface form of text and the higher-level structures of meaning, and LLMs get us part of the way there by exploiting priors in context. Usage-based theories can then provide the conceptual tools to take us the rest of the way, enabling a more systematic connection between linguistic form, meaning, and true human-like generalization.

In sum, LLMs offer a powerful but incomplete bridge between raw text and meaning. Their strengths lie in exploiting priors through context, but their limitations highlight the need for theoretical frameworks that go further. Usage-based approaches such as CxG provide exactly this. By combining the empirical reach of LLMs with the conceptual depth of usage-based theory, we can move toward a more systematic account of how form and meaning connect, and build systems that move towards human-like generalization.

8 Conclusions & Recommendations

The 'stochastic parrots' versus 'sparks of AGI' debate has become a roadblock to clarity in LLM performance and avenues to advance performance. This paper offers a more productive, middle-ground theory, providing a theoretically-grounded argument for context-directed extrapolation from training priors. The implications of this are significant: it provides a coherent explanation for the seemingly idiosyncratic and unpredictable strengths and weaknesses of LLMs, demystifying phenomena like hallucinations and, as we have detailed, clarifying their contradictory performance on CxG tasks.

Second, it suggests that meaningful improvements can be achieved not just through scale, but through better methods of directing this extrapolation via prompting and fine-tuning. This understanding demands that we re-evaluate how we improve language models. The prevailing paradigm, which chases unpredictable 'emergent' abilities by scaling up models and data, is not the only way forward. Our work suggests a more principled approach: focusing on the 'context' and 'priors' of the reasoning equation to achieve significant performance gains. This shift opens exciting new avenues for research beyond a simple reliance on scale. It points toward a more sustainable path to innovation, focused on augmenting models in novel ways, such as by equipping them with external memory.

Finally, and most urgently, our work demands a paradigm shift in how we evaluate LLMs. To genuinely measure a model's reasoning, we must move past benchmarks that might be tainted by training data or that only test for simple extrapolation. The goal should be to assess a model's ability to generalize and apply knowledge, not just to understand or remember it (in terms of Bloom's taxonomy (Bloom et al., 1956)). We therefore recommend a new focus on out-of-distribution evaluation, using grounded linguistic theory like CxG for language tasks. By testing models on examples that are grammatically valid but highly unlikely to be in the training data, such as formally identical but semantically distinct constructions, we can clearly distinguish between true generalization and mere pattern-matching.

Taken together, these recommendations call for a shift from chasing scale to building a linguistically principled science of evaluation and improvement, where CxG and related usage-based theories play a central role.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models.

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Irina Bigoulaeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. The inherent limits of pretrained llms: The unexpected convergence of instruction tuning and in-context learning capabilities.

Benjamin S Bloom et al. 1956. Taxonomy of. *Educational Objectives*.

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)*.

Claire Bonial, Julia Bonn, and Harish Tayyar Madabushi. 2025. Dancing with deer: A constructional perspective on mwes in the era of llms.

Claire Bonial and Harish Tayyar Madabushi. 2024a. Constructing understanding: on the constructional information encoded in large language models. *Language Resources and Evaluation*, pages 1–40.

Claire Bonial and Harish Tayyar Madabushi. 2024b. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

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. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

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

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Charles J. Fillmore, Russell Lee-Goldman, and Russell Rhomieux. 2012. The framenet constructicon. In Hans C. Boas and Ivan A. Sag, editors, *Sign-Based Construction Grammar*, number 193 in CSLI Lecture Notes, pages 309–372. CSLI Publications, Stanford, CA.

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

Adele Eva Goldberg. 1992. *Argument structure constructions*. University of California, Berkeley.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.

Steve Hanneke, Adam Tauman Kalai, Gautam Kamath, and Christos Tzamos. 2018. Actively avoiding nonsense in generative models. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 209–227. PMLR.

Thomas Hoffmann and Graeme Trousdale. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

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.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023a. Transformers as algorithms: generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023b. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR.

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?

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, Bangkok, Thailand. Association for Computational Linguistics.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2025. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models.

Jingcheng Niu, Subhabrata Dutta, Ahmed Elshabrawy, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. Illusion or algorithm? investigating memorization, emergence, and symbolic processing in in-context learning.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads.

Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025a. Unpacking let alone: Human-scale models generalize to a rare construction in form but not meaning. *arXiv preprint arXiv:2506.04408*.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025b. Assessing language comprehension in large language models using construction grammar.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023a. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, et al. 2023b. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs].

Harish Tayyar Madabushi, Taylor Pellegrin, and Claire Bonial. 2025a. Generative framenet: Scalable and adaptive frames for interpretable knowledge storage and retrieval for llms powered by llms. *Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning@ COLING 2025*, page 107.

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

Harish Tayyar Madabushi, Melissa Torgbi, and Claire Bonial. 2025b. Neither stochastic parroting nor agi: Llms solve tasks through context-directed extrapolation from training data priors.

Michael Tomasello. 2009. The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

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*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2024. Larger language models do in-context learning differently.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022a. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2024. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 3804–3811, Torino, Italia. ELRA and ICCL.

The potential of c2xg's unsupervised learning for metaphor extraction in African American novels

Kamal Abou Mikhael
University of British Columbia
Vancouver, Canada
kamalabm@student.ubc.ca

Abstract

This paper presents a pilot study of metaphors of motion in African American literary language (AALL) in two sub-corpora of novels published in 1920-1925 and 1926-1930. It assesses the effectiveness of Dunn's (2024) unsupervised learning approach to computational construction grammar (c2xg) as a basis for searching for constructional metaphors, a purpose beyond its original design as a grammar-learning tool. This method is chosen for its statistical orientation and employed without pre-trained models to minimize bias towards standard language; its output is also used to choose a target search term. Focusing on the verbal phrase 'come to', the study analyzes argument-structure constructions that instantiate conceptual metaphors, most prominently experiencer-as-theme (e.g., 'he came to know') and experiencer-as-goal (e.g., 'thoughts came to her'). The evaluation compares c2xg coverage against a manually annotated set of metaphors and examines the uniformity of metaphor types extracted. Results show that c2xg captures 52% and 63% of metaphoric constructions in the two sub-corpora, with variation in coverage and uniformity depending on the ambiguity of the construct. The study underscores the value of combining computational and manual analysis to obtain outcomes that are both informative and ethically aware when studying marginalized varieties of English.

1 Introduction

Developments in Construction Grammar (CxG), Conceptual Metaphor (CMT), Natural Language Processing (NLP), and the study of African American English (AAE) call for linguistic inquiry that goes beyond the vernacular (AAVE) and uses computational tools to explore the construal of the African American experience in metaphoric constructions.[1] Accordingly, this paper documents the initial phase of a longitudinal study of metaphors of motion in the literary language of African American novels published between 1920 (Harlem Renaissance) and 1975 (Black Arts). Here the data consists of novels from 1920-1925 and 1926-1930. Given that a marginalized variety should be studied without bias toward dominant varieties of English, Dunn's (2024) unsupervised learning approach to computational construction grammar (henceforth c2xg) is considered a suitable candidate for such a project because it can learn a CxG grammar without prior training or pre-existing biased models. However, in addition to the ethical criteria, the grammar it generates must be assessed for its effectiveness in searching for constructional metaphors, a purpose beyond its original design as a grammar-learning tool. The evaluation compares its coverage against a manually annotated set of metaphors and examines the uniformity of metaphor types extracted.

The paper focuses on argument structure constructions containing metaphoric usages of 'come to' whose meaning is distinguished by constraints on pairing and positioning of arguments. The results show that most such constructions in the corpus are experiencer-as-theme (e.g., 'he came to know/realize/think') and experiencer-as-goal (e.g., 'thought/suspicion/love came to her'), along with idioms such as 'come to think of it', and 'what is coming to you'.

2 Background

Study of AAE began with a focus on African American Vernacular English (Labov, 1972), but has come to include African American Standard English, African American Middle Class English,

[1] Some use the term *Language* instead of *English*, thus AAE and AAVE are also referred to as AAL and AAVL, respectively.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 202–207

African American Church Language, and various regional and demographic varieties (Bloomquist et al., 2015). The study of the language of African American literature has mainly focused on the representation of vernacular speech (Holton, 1984; Bailey, 1965; Williamson, 1970; Minnick, 2010; Green, 2002). Moreover, CMT based studies in African American literature have focused on a single work or a single author (Levinson, 2012; Mensah, 2011). This project focuses on the entirety of the language of literary works, a variety of AAE described as African American Literary Language (AALL). The language of novels is of interest because it evolved over time in two aspects: the representation of vernacular speech in dialogue, and the integration of vernacular forms into narration (Wideman, 1977). To approach AALL without assumptions, the study uses c2xg's unsupervised learning to discover metaphoric constructions that characterize AALL (Dunn, 2024).

3 Data

The data consists of the literary corpus and the output of the c2xg Python package. The corpus consists of novels from a list curated by the History of Black Writing Project (The Project on the History of Black Writing, 2024, 1987). For each time period (1920-1925 and 1926-1930) the works are narrowed down to ten based on the availability of the digital text and a strong element of realism which allows one to access metaphors that construe the African American experience.[2] The 1920-1925 sub-corpus contains 573,113 tokens and 1926-1930 contains 654,918. For each sub-corpus, a CxG grammar is generated using c2xg's Python implementation (Dunn, 2025). The CxG grammars are the generated from a single round of learning using the default parameters based on 500,000 words of each sub-corpus. Each grammar is a list of computationally derived constructions that can be augmented with examples from the corpus which are instantiations of the construction. These are referred to as *examples* in the c2xg documentation and they are henceforth referred to as c2xg examples to distinguish them from standard numbered linguistic examples listed in the paper. In this study c2xg is run to list all of the c2xg examples from which each construction are derived.

[2]This eliminates works of satire and historical fiction.

4 Methodology

Given the status and history of marginalization of the language being studied, the methodology prioritizes minimizing algorithm and human bias. To minimize algorithm bias, it uses the unsupervised learning of c2xg instead of language models and POS taggers that are skewed towards dominant varieties of English (Jørgensen et al., 2016; Hovy and Prabhumoye, 2021; Ziems et al., 2022) ; in addition, c2xg is run without any of its pre-trained models. To mitigate human bias, the target motion verb is chosen based on its frequency within the corpus and the metaphoric meanings it exhibits in the output of c2xg. The frequency is calculated from a word frequency list containing various inflections of 'come' (i.e., 'come', 'comes', 'came', 'coming', 'comin', and 'comin').

Although 'go' is more frequent than 'come' in the corpus, the latter is chosen because in the c2xg examples the verbal phrase 'come to' exhibits greater metaphoric variety in terms of argument structure. Metaphoric uses of 'go to' mainly consist of 'X going to Y' constructions in which subject X intends to take action Y (e.g., 'I'm going to quit', 'she was going to sleep'). On the other hand, 'come to' exists in a variety of metaphoric schematic (e.g., 'X comes to Y, Y=VP or NP: 'he came to love', 'he came to a decision') and idiomatic (e.g., 'come to think of it') constructions.

In order to assess the usefulness of c2xg for finding constructional metaphors two data sets are created. First, for each sub-corpus, an *evaluation set* is created using *verb-based* search (i.e., 'come to'). The results of the search are manually formatted to create a set of key-word-in-context (KWIC) entries where each 'come to' construction is annotated. Second, the evaluation sets are searched using corresponding c2xg examples (e.g., 'come to love') to create a pairs of search terms and matches (i.e., c2xg example and corresponding KWIC entry). The result is a *retrieved set* for each sub-corpus. These two sets are the basis for measuring *coverage* and *uniformity*.

4.1 Evaluation Set

The corpus is searched for 'come to' construct ('come to', 'came to', 'coming to', 'comin to', 'cominto', 'coming to', and 'comes to') to create a set of KWIC entries. Each 'X comes to Y' construction is delimited and marked according to its type: experiencer-as-theme, experiencer-

c2xg Example	KWIC Entry
['came', 'to', 'know']	Avey and my real relation to her, I thought I [came to know]+Verb.
['came', 'to', 'realize']	And although Peter [came to realize it]+Verb later it was many years before he told her so.

Table 1: c2xg examples and corresponding annotated KWIC entries.

as-goal, goal-as-physical-part (e.g., 'A determined look came to his face.'), other metaphor, and non-metaphor. Furthermore, the theme and goal in the construct are given semantic labels to distinguish them from other types of arguments (e.g., in 'come to love', the goal 'love' is labeled as 'Verb', and in 'peace came to her', the theme 'peace' is labeled as 'State').

4.2 Retrieved Set

The c2xg examples containing the various inflections of 'come to' are used to search the evaluation set. The search terms and matching results are paired to form the retrieved set. Table 1 shows a sample of entries; the '+' is shorthand annotation to mark experiencer-as-theme constructions.

4.3 Coverage and Uniformity

The results of the these two steps are used to evaluate *coverage* and *uniformity*. *Coverage* is the percentage of metaphoric constructions in the evaluation set that are found in the retrieved set ($metaphors\_retrieved/metaphors\_identified$). *Uniformity* is measured for c2xg examples that retrieve more than one result, and is the percentage of result sets that contain the same type of metaphor ($uniform/uniform + varied$). It is measured after the result set of each c2xg example is assigned a label. A result set consisting of one match is *single* and not part of the measure. Otherwise, if a result set has multiple matches which consist of the same metaphor type, it is *uniform*; otherwise, it is *varied*. These two measures indicate the usefulness of *c2xg* for locating metaphoric constructs.

5 Results and Analysis

Although the goal of this study is to assess coverage and uniformity, such a discussion is informed by an overview of the linguistic findings in the evaluation set. The first subsection gives an overview of the metaphoric constructions identified and annotated in the evaluation set. The second subsection provides an account of the coverage and uniformity.

5.1 Metaphoric Constructions

The main linguistic findings consist of the experiencer-as-theme and experiencer-as-goal schematic constructions that also include idioms. These constructions constitute the majority of the metaphors in the verb-based search results: 71.34% in 1920-1925 and 81.01% in 1926-1930.[3] In experiencer-as-theme, the experiencer is the theme because it is the subject of 'come to' and it can be a character (1a), group (1b), or general referent in the novel (1c). The experience it undergoes is construed with the conceptual metaphor CHANGE OF MENTAL STATE IS CHANGE OF LOCATION. The goal argument is realized through verbs featuring mental verbs or the verb 'be', and nominals denoting a state, event, or result.

(1) a. Peter came to realize it later
b. the cubs came to know him
c. handwriting all had come to know
d. he came to think it possible that
e. having come to understand
f. he had come to feel

The verbs convey cognition or perception such as 'realize', 'know', 'think', 'understand', and feel (1a-1f). The verb 'think' is also part of the idiom 'come/came to think of it' in which the experiencer can be implicit (2a) or explicit (2b). Other mental verbs include verbs of emotion such as 'love' (3a) and 'hate' (3b). Another notable verb is 'be' which introduces a state or result (4).

(2) a. Come to think of it they were ...
b. How'd you come to think of it?

[3]These results are calculated over example types (unique sequences of one or more tokens), as the same example instance may occur under multiple constructions, and some constructions are duplicates that yield identical sets of example instances.

(3) a. Just how I came to love her ...
b. We might come to hate each other

(4) a. you might come to be ashamed of me
b. puzzled by how they came to be there

Nominals of state or result in experiencer-as-theme include idiomatic (5a-5c) and compositional (5d-5f) constructs. In 'came/coming to himself' (5a-5b) the experiencer and the goal may seem to be the same, but this analysis considers 'himself' to refer to an ideal state of sound judgment that the experiencer had to arrive at. 'Came to' does not have an explicit goal, but the experiencer is understood to regain (arrive to) consciousness (5c).

(5) a. he came to himself
b. coming to himself
c. when he comes to there'll be no ...
d. came to the conclusion
e. she had come to the parting of the ways
f. what we are coming to

(6) a. thoughts of his condition came to her
b. suspicion had come to her
c. the desires ... which had come to her
d. the peace which had come to her

In experiencer-as-goal, the goal can be a character, part of a character (e.g., ears, mind), or thought (e.g., 'her reception of him'). The theme mostly consists of mental phenomena such as thoughts (6a), perceptions (6b), desires (6c), and psychological states 6d. The `c2xg` examples often do not contain the theme ('X' in 'X came to Y'). For example, the construct 'visions of Lida came to him' is extracted with the `c2xg` example 'Lida came to him'. `c2xg` examples containing a fully formed 'X came to him/her' extract non-metaphoric constructs such as 'she/he came to him/her'.

Although in general the goal in experiencer-as-goal refers to broad, external, and passive mental phenomena, there are a few instances where they refer to states or results which are the goal in experiencer-as-theme constructions. For example, the state 'senses' is the theme in (7a) whereas it is the goal in (7b). A similar example for the result 'decision' is found in (7c) and (7d).

(7) a. Then her senses came to her.
b. Just wait till you come to your senses!
c. A swift decision came to her.
d. before she could come to any decision

5.2 Coverage and Uniformity

`c2xg` example-based search has a total coverage of 51.79% and 62.5%. Table 2 shows a more detailed breakdown for each type of metaphor. It is evident that the individual coverage for each of these two constructions are not consistent across the corpora. Experiencer-as-theme has higher coverage in 1920-1925 whereas experiencer-as-goal's coverage is higher in 1926-1930. Certain metaphors are not extracted due to frequency of the search token sequence on the `c2xg` examples. For example, the phrase 'came to her' appears 13 times in 1920-1925, 7 of which are metaphors; in 1926-1930, it appears 42 times, 37 of which are metaphors. As a result, 'came to her' is represented in the grammar of 1920-1925 and not in 1926-1930.

Table 3 summarizes the measure of uniformity across metaphor types and non-metaphors. For the experiencer metaphors (-as-theme and -as-goal), uniform sets have almost double the uniformity in 1920-1925 compared to 1926-1930 (93% vs 42%). This is partially explained by the size of the result set: on average, a `c2xg` example that matches more than one result extracts 4.85 in 1920-1925, and 9 in 1926-1930. The `c2xg` examples that account for the majority of the varied results are incomplete phrases such as 'come to the' and 'come to her'. In the case of 'the', the type of metaphor is determined by what follows, resulting in experiencer-as-theme (8a), experiencer-as-goal (8b), and other metaphors (8c and 8d). In the case of 'her', the pronoun can be objective (9a) (experiencer-as-goal) or possessive (9b) (experiencer-as-theme), but possessive does not necessarily predict the type of metaphor (experiencer-as-goal).

(8) a. she ... came to the conclusion
b. sorrow ... come to the singers
c. she had come to the parting of the ways
d. the ... pessimist ... came to the front

(9) a. the thought had come to her
b. reluctant to come to her journey's end
c. phrases of thanks came to her mind

These preliminary observations of coverage and uniformity show that the evaluation of a tool like `c2xg` can inform how it is used and what is expected of it. In the case of coverage, one can expect `c2xg` to reduce the problem space by learning constructions of higher frequency, and a larger sample

Period	Source	Experiencer as Theme	Experiencer as Goal	Goal is Physical Part	Other Metaphor	Total
1920-1925	c2xg	34	24	0	9	69
	Corpus	56	56	6	39	157
	Coverage	61%	43%	0%	23%	43%
1926-1930	c2xg	21	59	3	8	91
	Corpus	48	80	5	25	158
	Coverage	44%	74%	60%	32%	58%

Table 2: Frequencies of 'come to' metaphors extracted by verb vs. c2xg examples.

Period	Source	Experiencer as Theme	Experiencer as Goal	Goal is Physical Part	Other Metaphor	Non-Metaphor
1920-1925	Single	28	6	0	2	45
	Uniform	13	4	0	3	23
	Varied	1	0	0	2	1
	Uniformity	93%	100%	0%	60%	96%
1926-1930	Single	15	3	0	3	79
	Uniform	5	2	0	0	32
	Varied	7	6	1	5	14
	Uniformity	42%	25%	0%	0%	70%

Table 3: Uniformity of metaphoricity in result sets extracted by c2xg examples.

size may ensure that key metaphors are not omitted. In the case of uniformity, c2xg examples consisting of incomplete phrases may reduce uniformity, but also may increase variety. Thus, the goals of the project would determine whether such phrases are used in the search process or whether they would need to be manually expanded in order to ensure greater uniformity in the results.

6 Conclusion

This paper described an ongoing-research project that is in its initial phases. It outlined a process for the use and evaluation of c2xg which was used to establish a statistical basis to identify potential metaphoric constructions. The metaphor analysis and argument structure analysis were fully manual. However, the process of extracting potential metaphors was done both computationally and manually which allowed for c2xg to be evaluated for uses beyond its intended purposes.

The main contribution of this study was insights on how the output of c2xg affects the extraction of metaphoric constructions so that it can be used in a manner that serves a project's objectives. Still, the usefulness of its data was not exhausted. Additional c2xg runs and rounds of learning are necessary, and there remains analysis of the computationally derived constructions and their translation into human-legible argument structure CxNs.

In the context of studying marginalized varieties of English, the unsupervised learning approach of c2xg presents a relatively safe start. However, given the presence of General American English (GAE) in the data encountered, at least for the period observed (1920-1930), there may be potential for the use of existing English NLP tools whose output can be monitored and evaluated so that they can be modified or that their output can be used in a manner that is less biased. The mindful and vigilant interplay between the computational and manual analysis of constructions and metaphor analysis is key for obtaining outcomes that are informative and ethically aware.

Acknowledgments

I would like to acknowledge Jonathan Dunn whose approach to computational CxG and the Python package he implemented were central in this paper along with his generous feedback and support. I also would like to acknowledge the feedback of my advisor Dr. Elise Stickles at the University of British Columbia.

References

Beryl Loftman Bailey. 1965. Toward a new perspective in Negro English dialectology. *American Speech*, 40(3):171–177.

Jennifer Bloomquist, Lisa J. Green, and Sonja L. Lanehart, editors. 2015. *The Oxford Handbook of African American Language*. Oxford University Press.

Jonathan Dunn. 2024. *Computational Construction Grammar: A Usage-Based Approach*. Elements in Cognitive Linguistics. Cambridge University Press.

Jonathan Dunn. 2025. jonathandunn/c2xg. Original-date: 2016-05-22T21:03:06Z.

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Sylvia Wallace Holton. 1984. *Down Home and Uptown: The Representation of Black Speech in American Fiction*. Fairleigh Dickinson University Press ; Associated University Presses, Rutherford : London.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California. Association for Computational Linguistics.

William Labov. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press, Philadelphia.

Julian Levinson. 2012. All the metaphors you are: Conceptual mappings of bebop in James Baldwin's Sonny's blues and Jack Kerouac's On the road. *Jazz research journal*, 6(1):69–87.

Eric Opoku Mensah. 2011. The metaphor: A rhetorical tool in some selected speeches of Martin Luther King, Jr. and Kwame Nkrumah. *Language in India*, 11(4):155–172. Publisher: Language in India.

Lisa Cohen Minnick. 2010. *Dialect and Dichotomy: Literary Representations of African American Speech*. The University of Alabama Press, Tuscaloosa.

The Project on the History of Black Writing. 1987. History of Black writing novel collections.

The Project on the History of Black Writing. 2024. History of Black writing (hbw) corpus.

John Wideman. 1977. Defining the Black Voice in Fiction. *Black American Literature Forum*, 11(3):79–82.

Juanita V. Williamson. 1970. Selected features of speech: Black and White. *CLA Journal*, 13(4):420–433.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding Dialect Disparity in NLU. ArXiv:2204.03031 [cs].

Author Index