Annotation Scheme for English Argument Structure Constructions Treebank

Hakyung Sung and **Kristopher Kyle**
Learner Corpus Research and Applied Data Science Lab
Department of Linguistics
University of Oregon
hsung,kkyle2@uoregon.edu

Abstract

We introduce a detailed annotation scheme for argument structure constructions (ASCs) along with a manually annotated ASC treebank. This treebank encompasses 10,204 sentences from both first (5,936) and second language English datasets (1,948 for written; 2,320 for spoken). We detail the annotation process and evaluate inter-annotation agreement for overall and each ASC category.

1 Introduction

The usage-based constructionist approach posits that language evolves from form-meaning pairings within linguistic structures. As language learners are exposed to a variety of inputs and outputs, they develop constructions of varying complexities (Goldberg, 2013). Among these, argument structure constructions (ASCs) represent core sentence meanings and are essential communication tools (Goldberg, 1995, 2003, 2009, 2013; Rappaport Hovav and Levin, 1998; O'Connor and Kay, 2003).

In cognitive linguistics, ASCs have played a central role in shaping cognition and language learning. Language learners categorize experience into structured frames. During this process, ASCs provide a conceptual framework, especially during the initial period of interaction with their surroundings. For example, with a transitive ASC, language learners can articulate a "conceptual archetype" related to situations in which a subject affects an object (e.g., *I kick the ball*) (Langacker, 1987). The significance of ASCs has been further underscored by research for both first (L1; Cameron-Faulkner et al., 2003; Goldberg, 2013, 2019; Ninio, 1999) and second language (L2; Ellis, 2002; Ellis and Larsen-Freeman, 2009; Ellis and Ferreira-Junior, 2009; Kyle and Crossley, 2017) learning. Recently, computational linguists have also been interested in investigating how well constructional information is learned and induced by large language models (Dunn, 2017; Li et al., 2022; Madabushi et al., 2020, 2023; Weissweiler et al., 2023).

In this paper, we introduce an ASC annotation procedure, its accompanying treebank (22,069 ASC tokens), and a supplementary annotation manual (32 pages). Our work includes both L1 and L2 English datasets, which serves two purposes: to augment the first version of the ASC treebank (Kyle and Sung, 2023), and to enable a comprehensive analysis across different linguistic contexts. The inclusion of L2 data is particularly helpful for researchers in the field of L2 acquisition (e.g., Kyle et al., 2021) and for developers of automatic writing evaluation systems used in language proficiency assessments (e.g., Yancey et al., 2023). Previous findings have suggested that the accuracy of part-of-speech and dependency tagging models on L2 datasets improves when these models improves on L2 datasets when L1 training data is supplemented with even relatively small amounts of L2 data (e.g., Kyle et al., 2022; Sung and Shin, 2023).

This dataset is expected to be instrumental in the evaluation and enhancement of tools for linguistic feature analysis, particularly for ASC-related statistical measures, such as the association strength between verbs and ASCs (Gries and Ellis, 2015; Kyle, 2016). Additionally, the dataset would offer opportunities for investigating how large language models process ASCs within both L1 and L2 contexts.

2 Related Work

Research into ASC annotation has been limited. Some pioneering efforts have focused on specific ASC types, such as a caused-motion ASC (Hwang, 2014) or ASCs with light verbs (e.g., *take, make*) (Bonial, 2014). Recently, Kyle and Sung (2023) developed a publicly available silver-annotated treebank of ASCs based on previous related projects such as PropBank (Palmer et al., 2005), FrameNet (Fillmore et al., 2003), VerbNet (Schuler, 2005)

Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII), pages 12–18

and Universal PropBank (UP) for multilingual semantic role labeling (Akbik et al., 2015). Using a semi-automatic approach, they extracted large-grained semantic role frames from PropBank (e.g., *ARG0-Verbsense-ARG1*) then refined them (e.g., *agent-Verbsense-theme*) using PropBank mapping protocols with FrameNet and VerbNet databases. They then manually assigned each semantic role frame to an ASC (e.g., *agent-Verbsense-theme* → Transitive simple ASC). After resolving ambiguities and conducting spot-checks for errors, they categorized 94.1% of the ASCs (n = 26,437)[1] in the English Web treebank (EWT) (Bies et al., 2012; Silveira et al., 2014). The reliability of the semi-automatic annotation was preliminarily evaluated using a small sample of manually annotated sentences (n = 100) from the treebank. The results suggested that the semi-automatic approach was reasonable, but far from perfect (simple agreement rate = 92.1%, *kappa* = .884). While this approach presents an initial effort in building the silver-standard ASC treebank, the accuracy of their semi-automatic approach is still unclear.

Beyond annotations, researchers have explored the relationship between ASCs and language learning or understanding (recently processed by large language models) in various ways. These include psycholinguistic experiments (Bencini and Goldberg, 2000; Li et al., 2022), manual extractions based on researchers' judgments (Ellis and Ferreira-Junior, 2009), automatic extractions that leverage the syntactic forms of the ASCs (using automatic dependency annotations) (Kyle and Crossley, 2017; Hwang and Kim, 2023). Furthermore, some studies delved into broader linguistic constructions, identifying them through unsupervised approach that leverage statistical association measures (Dunn, 2017; Madabushi et al., 2020).

3 Annotation Scheme

In the development of the annotation scheme, we first address two important levels of linguistic annotation (Nivre et al., 2004). The first level is selecting the annotation unit, and the second level is choosing the ASC categories to be employed during the annotation process.

[1]For a detailed representation of ASCs in the treebank, including the frequencies and the most frequent verbs of each ASC, see Kyle and Sung, 2023, p. 54, Table 1).

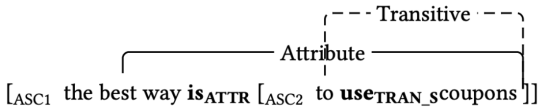

Figure 1: Example of annotation unit. In visual representations, a solid line denotes an ASC span in a finite clause form, while a dashed line indicates an ASC span in a non-finite form.

3.1 Analysis unit and basic principles

The unit of analysis is centered on the ASCs, which built upon arguments interpreted by specific semantic roles, specifying *Who did What to Whom* (e.g., agent, beneficiary, goal, theme)[2] in the given context (Goldberg, 1995).

Verbs serve as annotation targets due to their interactions with adjacent arguments (Fillmore, 1968; Goldberg, 1995). While early studies on ASCs (e.g., Goldberg, 1995) predominantly discussed finite clauses as ASCs, we encompassed both finite clauses and/or non-finite clauses. This comprehensive approach ensures every meaning constrained by an ASC is captured. As a result, some sentences are parsed into multiple layers which include both finite and non-finite clauses, with each layer presenting a distinct ASC (Figure 1).

Word order is considered a separate construction layer. This means we treated different word orders for the same ASC type (due to pragmatic reasons like framing a question) as identical[3]. This approach aligns with the understanding that ASCs do not dictate specific word orders (Goldberg, 2019, p. 39).

3.2 ASC categories

We adopted nine ASC categories. Meanwhile, we recognize the flexibility in ASC selection (Kyle and Sung, 2023) and the varying complexity levels in ASC representation (e.g., focusing solely on light verb constructions; Bonial, 2014).

Initially, our focus was on seven primary ASCs, predominantly characterized by active voice as often outlined in earlier studies (Bencini and Goldberg, 2000; Ellis and Ferreira-Junior, 2009; Goldberg, 1995; Hwang and Kim, 2023): (1) intransitive simple (e.g., ***worked*** *a lot*); (2) intransitive motion

[2]For a comprehensive set of recognized semantic role labels, we refer to Palmer et al., 2011, p. 4.

[3]For example, we may tag two expressions with different syntactic frames *they* ***wanted*** *the thing* and *the thing which they* ***wanted***, as TRAN_S, even though the latter has an inverted object (*the thing*) due to its relative clause structure.

(e.g., ***went*** *out the door*); (3) intransitive resultative (e.g., ***got*** *healthier*); (4) transitive simple (e.g., ***met my friends***); (5) ditransitive (e.g., ***told*** *them my plan*); (6) caused-motion (e.g., ***put*** *the date on the calendar*); (7) transitive resultative (e.g., ***made*** *me happy*).

In our analysis, we choose to expand our category beyond the active voice to comprehensively account for the ASC types appearing in the data. As a result, we augmented our category with two additional constructions: (8) attributive (e.g., ***be*** *the first*) and (9) passive. Furthermore, we classified the sentences with middle voice (e.g., *the pot* ***broke*** *on the floor*) under our (1) intransitive simple category because of its syntactic resemblance. While our categorization aligns with the goal of this study, future research may adopt finer-grained approaches to sub-categorize each voice tag. Different voices such as active, middle, passive have the potential to be considered distinct layers (Goldberg, 2013, p. 455). These voices may play an important role in shaping syntactic/semantic frames, altering dynamics of agency and focus within sentences[4].

Table 1 illustrates the nine ASC types included in this study, with the most prototypical syntactic[5] and semantic frame representations. Note that these frames were not utilized as strict criteria for annotating ASCs. Instead, they represent typical or frequently encountered patterns observed during our annotation process and in comparisons with the silver ASC treebank. We have compiled a more comprehensive set of syntactic and semantic frames in our annotation manual, which served as a detailed reference for the annotators.

4 Datasets

We utilized L1 and L2 English datasets that were publicly and freely available and are part of Universal Dependency (UD) projects. Both datasets use the CoNLL-U format, which presents sentences vertically with columns detailing morphological and syntactic attributes. We annotated ASC tags into the 10th MISC column. Due to limited annotation capacity, we randomly selected a subset of sentences from L1 and L2 written datasets.

4.1 L1 dataset

Following the related work (Kyle and Sung, 2023), we continued to use the English portion of the UP project, which combined the UD version of the EWT with semantic role labels based on the PropBank annotation scheme. The original EWT corpus contains sentences sampled from five web registers: blogs, newsgroups, emails, reviews, and Yahoo Answers. Within the EWT's total of 16,621 sentences, we manually tagged 5,936 sentences (104,640 word tokens).

4.2 L2 dataset

We used both L2 written (ESL-WR) and spoken (ESL-SP) corpora to cover a broader range of L2 English registers.

4.2.1 L2 written

The ESL-WR dataset (Berzak et al., 2016) is sourced from the CLC FCE dataset (Yannakoudakis et al., 2011), containing written answers from the Cambridge English exams across five registers (letter, report, article, composition, and short story). Extracts in the ESL dataset came from upper-intermediate English learners representing 10 different native languages. Of the 5,124 sentences in the original corpus, we manually tagged 1,948 sentences (37,055 word tokens).

4.2.2 L2 spoken

The ESL-SP dataset (Kyle et al., 2022) is derived from sentences randomly sampled from the NICT JLE (Izumi et al., 2004) corpus. This corpus features transcriptions from oral proficiency interviews that contained utterances produced by Japanese English learners. We manually tagged all 2,320 sentences (21,312 word tokens) that have annotations with syntactic dependency relation tags in the corpus.

5 Annotation

The annotation project spanned about 8 months, from April 2023 to November 2023.

[4]For example, when a sentence *a car* ***hit*** *the man* is transformed into *the man was* ***hit*** *by a car*, the arguments and their semantic roles remain consistent (*car*-agent; *man*-theme), but only the focus within the statement is shifted. This allows for ASCs from active voice categories, which have *agent* and *theme* arguments (i.e., transitive simple, ditransitive, cause-motion, transitive resultative), may have respective passive counterparts and could potentially be grouped into separate passive categories. However, in our current categorization, we have grouped them all under the PASSIVE tag. We acknowledge that this is a methodological decision and that exploring distinct passive categories for different ASC types could be a valuable direction for future research.

[5]Syntactic frames are described by syntactic dependency tags, drawing upon the tags from the UD project: https://universaldependencies.org/u/dep/index.html.

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

Table 1: ASCs representation in treebank

5.1 Annotator training

We enlisted and trained six undergraduate Linguistics majors, all of whom were native English speakers and had previously undertaken advanced courses in functional English syntax. These annotators participated in three structured 1-hour training sessions.

During the initial training session, the annotators were introduced to the overarching theoretical background and objectives, as well as their expected roles. They were also familiarized with the data format, tagging schemes, and the procedures for accessing and storing data within a shared folder. For the annotation process, we utilized standard text editors (e.g., *BBEdit*) and/or spreadsheet software (e.g., *Microsoft Excel*), which provided a straightforward interface for manual tagging. In the subsequent two sessions, they engaged in hands-on practice with sample sentences, individually tagging items and discussed any challenges they encountered. Feedback was provided after each exercise.

5.2 Annotation process and review

After training, the annotators began remote work, uploading files to a monitored shared folder. They also had access to a thoroughly documented webpage, which was updated regularly by the researchers, and a Discord server for discussing challenges. Based on these discussions, the tagging guidelines were frequently updated and eventually formed the basis of the annotation manual.

While the original L1/L2 datasets, sourced from the UD project, provided some syntactic cues (e.g., *nsubj, root*) our annotators were not strictly bound to these existing annotations. In other words, annotators had flexibility to rely on the matching patterns between UD tags and ASC categories when they deemed these reliable. They were also encouraged to ignore the UD tags in cases in which the UD syntactic frame did not align with the ASC's semantic arguments.

Each ASC token was primarily tagged by two randomly assigned annotators through a blind review process. If there was disagreement between the two annotators, a third annotation was done, either by one of the undergraduate annotators or one of the researchers. If disagreement persisted, a fourth annotator, typically one of the researchers, tagged the case.

5.3 Annotation summary and evaluation

Inter-annotator agreement during the first round of annotation was reasonable[6] (exact agreement = 85.7%; *kappa* = .801; Landis and Koch, 1977). Table 2 provides the number of annotated ASCs in each dataset. For a visual representation of annotator agreement on each ASC tag, refer to the confusion matrix depicted in Figure 2.

Tag	EWT	ESL_WR	ESL_SP
INTRAN_S	1,395	662	525
INTRAN_MOT	607	250	240
INTRAN_RES	213	44	23
TRAN_S	6,094	2,488	1,385
DITRAN	285	160	37
CAUS_MOT	766	87	53
TRAN_RES	763	76	16
ATTR	2,539	1,289	760
PASSIVE	1,058	224	50
Total	**13,720**	**5,260**	**3,089**

Table 2: ASCs distribution in treebank

[6]These figures are slightly lower when misspelled and missed tags are included (exact agreement = 82.5%, *kappa* = .759). These are not represented in the confusion matrix in Figure 2.

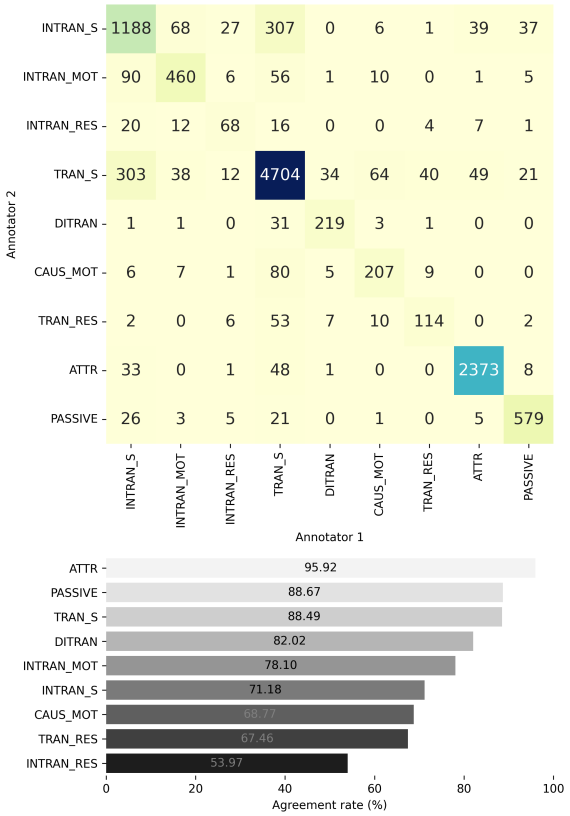

Figure 2: Evaluation of inter-annotator agreement

Semi-automatic annotations from the silver ASC treebank (11,245 ASCs) were also evaluated against the gold standard. By-tag evaluation scores are outlined in Table 3.

Tag	Count	P	R	F1
INTRAN_S	1,014	.895	.787	.837
INTRAN_MOT	353	.758	.691	.723
INTRAN_RES	124	.861	.649	.740
TRAN_S	5,195	.919	.909	.914
DITRAN	206	.851	.789	.819
CAUS_MOT	517	.807	.776	.791
TRAN_RES	673	.707	.917	.798
ATTR	2,445	.982	.972	.977
PASSIVE	718	.924	.948	.936

Table 3: Evaluation of semi-automatic annotations

5.4 Annotation manual overview

In categorizing ASCs, making informed choices is important, as often seen in other annotation projects (Gerdes and Kahane, 2016). Our annotation manual details the "what" and "why" behind the annotators' choices, providing various exemplar sentences from the ASC treebank. These highlight various verb usages in each ASC, associated with syntactic frames (based on dependency tags) and, for L1 data, semantic frames (based on semantic role labels). The manual also addresses challenges in annotating ASC tags, such as ambiguous cases and complexities arise from verb-ASC combinations. Through the manual, we seek to ensure transparency and consistency in both current and future endeavors.

6 Conclusion

In this study, we introduced a manually annotated ASC treebank, complemented by a detailed annotation procedure. Access to the ASC treebank and the accompanying annotation manual is available here: https://osf.io/v75qu/?view_only=410a39910fd1438bbf1ef0368ce51c3b.

Limitations

The study has limitations that should be addressed in future work. First, the current ASC treebank is confined to English, limiting the applicability of findings to other languages. Second, the tagging scheme could be refined to capture subtle language nuances (e.g., PASSIVE tag). Third, expanding the dataset to include a broader range of genres and registers would improve the understanding of ASC patterns.

Ethics Statement

The annotated dataset presented in this work utilized publicly available datasets from previous studies, including the silver-standard ASC treebank (Kyle and Sung, 2023), ESL-WR (Berzak et al., 2016), ESL-SP (Kyle et al., 2022), and Universal PropBank (Akbik et al., 2015).

Acknowledgements

The authors would like to express their gratitude to the annotators—Hana Dussan, Bella Graham, Marianne Hatley, Augustus Paddock, Holly Renshaw, and Ryan Walker (listed in alphabetical order)—for their contributions to the manual annotations and the valuable discussions that have improved the ASC tagging guidelines.

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings*

of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 397–407.

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.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner english. *arXiv preprint arXiv:1605.04278*.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Claire Nicole Bonial. 2014. *Take a look at this! Form, function and productivity of English light verb constructions*. Ph.D. thesis, University of Colorado at Boulder.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive science*, 27(6):843–873.

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

Nick C Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2):143–188.

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. 1968. Lexical entries for verbs. *Foundations of language*, pages 373–393.

Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Background to framenet. *International journal of lexicography*, 16(3):235–250.

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *LAW X (2016) The 10th Linguistic Annotation Workshop: 131*.

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

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

Adele E. Goldberg. 2009. Constructions work. *Cognitive Linguistics*, 20(1):201–224.

Adele E Goldberg. 2013. The emergence of the semantics of argument structure constructions. In *The emergence of language*, pages 215–230. Psychology Press.

Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

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

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.

Jena D Hwang. 2014. *Identification and representation of caused motion constructions*. Ph.D. thesis, University of Colorado at Boulder.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The nict jle corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, 12(2):119–125.

K. 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.

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

Kristopher Kyle, Scott Crossley, and Marjolijn Verspoor. 2021. Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, 43(4):781–812.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language english. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45.

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.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

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

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. *arXiv preprint arXiv:2202.12246*.

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

Harish Tayyar Madabushi, Laurence Romain, Petar Milin, and Dagmar Divjak. 2023. Construction grammar and language models. *arXiv preprint arXiv:2308.13315*.

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

Joakim Nivre, Koenraad De Smedt, and Martin Volk. 2004. Treebanking in northern europe: A white paper. *Nordisk Sprogteknologi*, pages 2000–2004.

Mary Catherine O'Connor and Paul Kay. 2003. Regularity and idiomaticity in grammatical constructions: The case of let alone. In *The new psychology of language*, pages 243–270. Psychology Press.

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

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2011. *Semantic role labeling*. Morgan & Claypool Publishers.

Malka Rappaport Hovav and Beth Levin. 1998. Building verb meanings. *The projection of arguments: Lexical and compositional factors*, pages 97–134.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A gold standard dependency corpus for english. In *LREC*, pages 2897–2904. Citeseer.

Hakyung Sung and Gyu-Ho Shin. 2023. Diversifying language models for lesser-studied languages and language-usage contexts: A case of second language Korean. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11461–11473, Singapore. 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. *arXiv preprint arXiv:2302.02178*.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.