

Proto-Role Semantic Role Labeling for Khmer: A New Approach for Argument Classification

Thach Thi Omnara

College of Southern Khmer Language, Culture, Arts and Humanities Tra Vinh University

DOI: <https://doi.org/10.47772/IJRISS.2026.100500341>

Received: 06 May 2026; Accepted: 11 May 2026; Published: 01 June 2026

ABSTRACT

This study examines the applicability of Dowty's (1991) Proto-role theory to Khmer, a morphologically impoverished and under-resourced language, with the aim of exploring its potential contribution to semantic role labeling (SRL) in languages lacking overt case marking and agreement. Based on a manually annotated dataset of 81 Khmer sentences drawn from literary sources, the study analyzes how Proto-Agent and Proto-Patient entailments are distributed across diverse constructions, including serial verb constructions, passive clauses, psychological predicates, and pro-drop structures.

The findings suggest that Proto-role entailments can provide a flexible and empirically useful framework for argument classification in Khmer, even in the absence of explicit morphosyntactic cues. The study aims to make three limited but targeted contributions: (i) it presents the first systematic Proto-role analysis for Khmer, (ii) it offers empirical observations indicating that Proto-role mappings remain largely stable across complex predicate structures, and (iii) it proposes a pilot annotation strategy that may inform the future development of Khmer semantic role resources for computational applications. While exploratory in scope, these results point to the potential of Proto-role-based SRL as a foundation for semantic annotation in under-resourced languages.

Keywords: Proto-role theory; semantic role labeling; argument structure; Khmer linguistics; under-resourced languages

INTRODUCTION

Semantic Role Labeling (SRL) has become an important task in natural language Semantic Role Labeling (SRL) aims to identify the semantic functions of sentence constituents relative to a predicate. Dowty's (1991) Proto-role theory reframes traditional theta roles as clusters of semantic entailments, with Proto-Agent and Proto-Patient categories defined by graded properties rather than fixed labels. Although Proto-role approaches have been discussed for several well-studied languages, their applicability to Southeast Asian languages such as Khmer remains underexplored. Khmer poses well-known challenges for SRL because grammatical relations are often weakly signaled morphologically and must be inferred from predicate semantics and context. SRL frameworks that rely heavily on case marking, agreement, or rigid argument positions therefore transfer only imperfectly to Khmer. This study explores whether an entailment-based Proto-role framework can capture recurrent patterns of argument realization in Khmer across a small but diverse dataset. This paper makes three limited but targeted contributions. First, it presents a systematic pilot application of Proto-role analysis to Khmer based on manually annotated data. Second, it examines whether recurrent Proto-role entailment patterns can be observed in constructions that complicate argument identification, including serial verb constructions, passives, and pro-drop contexts. Third, it outlines a preliminary annotation procedure that may serve as a conceptual foundation for future Proto-role-oriented semantic resources for Khmer. Rather than proposing a finalized semantic role labeling framework for Khmer, the present study should be understood as a feasibility-oriented exploratory investigation. Its primary aim is to evaluate whether Proto-role entailments can provide a sufficiently stable analytical basis for semantic role interpretation in a morphologically impoverished language where grammatical relations are often weakly signaled morphosyntactically.

Research Questions

This study addresses the following questions:

1. To what extent do Proto-Agent and Proto-Patient entailments account for argument realization in Khmer across different constructions?
2. How are Proto-role properties expressed in contexts with limited overt morphosyntactic role marking (e.g., serial verb constructions and passives)?
3. What implications does a Proto-role analysis have for future Khmer SRL resources and modeling?

METHODOLOGY

Data Selection

The dataset contains 81 Khmer sentences from literary texts (notably Pka Sropon by Nou Hach) and contemporary journalism (Khmer Times). Sentences were selected to cover a range of clause types relevant to Proto-role analysis, including transitive/intransitive clauses, serial verb constructions, passives, statives, and psychological predicates. Because this is a pilot study, selection prioritized semantic transparency and constructional diversity rather than frequency-based representativeness.

Annotation Procedure

Each sentence was annotated in three steps. First, the main predicate(s) were identified; serial verb constructions were treated as predicate complexes for Proto-role assessment. Second, core arguments were identified based on semantic participation, with adjuncts excluded from Proto-role assignment. Third, each argument was evaluated against a Proto-role entailment checklist (volition, sentience, causation, movement, affectedness, and change of state). Roles were assigned by comparing the relative concentration and salience of Proto-Agent versus Proto-Patient entailments. In clauses lacking a clearly affected participant, only a Proto-Agent classification was assigned.

Reliability

Annotation was performed by a single annotator, and inter-annotator agreement was therefore not measured. The results should consequently be interpreted as exploratory rather than statistically validated. In addition, Proto-role assignment inherently involves gradient semantic judgments, particularly in constructions involving psychological predicates, abstract affectedness, and serial predicate structures. Although the annotation procedure employed an explicit entailment checklist in order to improve internal consistency, some degree of interpretive flexibility remains unavoidable. Future work should expand the dataset and incorporate multiple annotators in order to evaluate annotation reliability quantitatively and to refine Khmer-specific Proto-role annotation guidelines.

Scope and Limitations

This study is limited to sentence-level Proto-role analysis and does not attempt discourse-level annotation, large-scale corpus modeling, or computational implementation. Because the project was designed as an exploratory pilot investigation, the dataset remains relatively small and was selected primarily to maximize constructional diversity rather than statistical representativeness.

The corpus is restricted to literary and journalistic Khmer and does not yet include conversational interaction, spontaneous spoken discourse, or digital communication data such as online messaging and social media text. As a result, the present findings should not be interpreted as fully representative of naturally occurring Khmer usage across all communicative settings.

A further limitation concerns the annotation procedure itself. Proto-role assignment relies partly on semantic interpretation, particularly in cases involving psychological predicates, metaphorical affectedness, and serial predicate constructions. While the use of an explicit entailment checklist helped maintain analytical consistency, the absence of multiple annotators limits the extent to which annotation reliability can presently be evaluated.

In addition, the study does not directly compare Proto-role analysis with alternative semantic role labeling frameworks such as PropBank or FrameNet through quantitative evaluation or parallel annotation experiments. The goal of the present study is therefore not to demonstrate the superiority of Proto-role theory, but rather to examine its feasibility and descriptive usefulness as an initial semantic annotation framework for Khmer.

Despite these limitations, the study provides preliminary evidence that entailment-based role analysis can capture recurrent semantic tendencies in Khmer argument structure and may offer a useful starting point for future SRL-oriented resource development.

LITERATURE REVIEW

Proto-role theory and analytic languages Dowty's (1991) Proto-role theory marks a fundamental shift from traditional theta-role frameworks by replacing discrete role inventories with clusters of semantic entailments that tend to characterize Agent-like and Patient-like arguments. Rather than assigning fixed labels, argument roles are determined by graded properties such as volitionality, sentience, causation, and self-initiated movement on the one hand, and affectedness, change of state, lack of control, and incremental themehood on the other.

This entailment-based perspective is particularly well suited to analytic languages, where morphological case marking and agreement provide limited guidance for role identification. In such languages, semantic role assignment must rely more heavily on lexical semantics and event structure than on surface morphosyntax. Subsequent work has shown that proto-role entailments interact systematically with verb classes and argument alternations, especially in constructions where traditional thematic roles become unstable or underspecified (Levin, 1993).

Methodologically, the literature suggests that Proto-role annotation is most effective when implemented through explicit entailment coding followed by a comparative decision rule that determines Proto-Agent or Proto-Patient status based on the relative dominance of entailments. Importantly, gradient and ambiguous cases—particularly those involving cognition and psychological predicates—are treated as theoretically motivated rather than as annotation noise (Dowty, 1991).

SRL annotation paradigms: PropBank, FrameNet, and Proto-roles

Two major paradigms dominate semantic role labeling research: PropBank and FrameNet. PropBank provides a large-scale, predicate-specific annotation scheme in which core arguments (Arg0–Arg5) and adjuncts are labeled for each predicate, enabling robust supervised modeling (Palmer et al., 2005). However, its reliance on predicate-specific rolesets and partial dependence on syntactic configurations pose challenges for cross-linguistic transfer, particularly for languages such as Khmer whose argument realization strategies diverge from English-centric assumptions.

FrameNet, by contrast, adopts a frame-semantic approach that models meaning through frames and frame elements grounded in corpus evidence (Baker et al., 1998). While semantically richer, FrameNet requires substantial annotation effort and extensive lexical coverage, which can be prohibitive for under-resourced languages.

Within this landscape, Proto-role entailments can be viewed as an intermediate semantic layer: less predicate-specific than PropBank and less lexicon-intensive than FrameNet. Prior work suggests that Proto-roles are especially useful in pilot studies and early-stage corpus development, where consistent cross-predicate semantic coding is needed before large-scale annotation becomes feasible (Dowty, 1991). As such, Proto-role labels may function either as a stand-alone analytical framework or as a feature layer supporting later PropBank- or FrameNet-style SRL.

Methodological Developments in SRL

Early shared tasks, particularly CoNLL-2005, played a central role in standardizing SRL evaluation and task decomposition, typically distinguishing argument identification from role classification (Carreras & Màrquez, 2005). Subsequent research demonstrated that syntactic parsing and global inference substantially improve SRL performance by enforcing structural consistency across role assignments (Punyakanok et al., 2008).

However, syntax-dependent approaches presuppose reliable parsing, which is often unavailable for under-resourced languages. Dependency-based SRL has therefore been proposed as a more practical alternative, integrating role labeling directly with dependency structures (Johansson & Nugues, 2008). This line of work is particularly relevant for Khmer, where dependency representations may ultimately prove more robust than phrase-structure analyses.

More recently, neural SRL models have achieved strong results in high-resource settings, including architectures that minimize explicit reliance on syntactic input (He et al., 2017). Nevertheless, these models typically require large annotated corpora and reliable segmentation, both of which remain challenging for Khmer. Taken together, the SRL literature reveals a methodological gap: most existing approaches assume either substantial labeled data or stable syntactic scaffolding, conditions that are not yet met for Khmer at scale.

Khmer Syntax, Typology, and Implications for SRL

Typological research has long identified serial verb constructions (SVCs) as a central mechanism for event representation in many languages, characterized by multiple verbs sharing arguments within a single clause (Aikhenvald & Dixon, 2006). Such constructions pose well-known challenges for argument identification and role assignment. Khmer-specific studies have documented the structural diversity of SVCs and related verbal constructions, providing valuable descriptive insights into argument sharing and event structure (Pearson, 2013; Self, 2014).

With respect to passivization, Kratzer's (1996) proposal that external arguments are introduced by functional structure rather than inherent to verbs offers a theoretical basis for separating syntactic position from semantic role. This distinction is especially relevant for Khmer passives, where Proto-Agent and Proto-Patient entailments may remain stable despite changes in grammatical subjecthood.

While existing Khmer scholarship offers detailed descriptions of constructions such as SVCs and passives, it has rarely operationalized these phenomena within an SRL-oriented annotation framework. In particular, there is little work that formalizes how argument roles in Khmer can be encoded in a way that is both semantically grounded and computationally extensible.

Khmer NLP Infrastructure and Research Gap

Recent advances in Khmer NLP provide a partial foundation for future SRL development. The Asian Language Treebank (ALT) project offers standardized tokenization and part-of-speech annotation guidelines, addressing a critical challenge posed by Khmer's lack of whitespace word boundaries (NICT, 2016). Additional studies on Khmer tokenization and POS tagging further document the practical constraints involved in processing Khmer text and help contextualize annotation decisions.

Nevertheless, existing resources remain largely focused on lower-level linguistic annotation and do not yet incorporate semantic role information. This gap motivates the present study, which explores Proto-role entailments as a lightweight yet theoretically grounded approach to semantic role annotation. By aligning Proto-role analysis with existing tokenization and POS standards, this work outlines a plausible pathway toward future dependency-based or neural SRL systems for Khmer.

Theoretical Background: Dowty's Proto-Role

Theory Traditional approaches to thematic roles typically assume a fixed inventory of discrete roles such as Agent, Patient, Experiencer, or Theme, which are assigned to arguments based on relatively rigid syntactic or lexical criteria. While such frameworks have proven useful for languages with rich morphosyntactic marking, they encounter difficulties when applied to languages in which argument realization is flexible and morphologically underspecified. In response to these limitations, Dowty (1991) proposed the theory of Proto-Roles, offering a reconceptualization of thematic roles grounded in semantic entailments rather than categorical labels.

At the core of Dowty's proposal is the distinction between Proto-Agent and Proto-Patient, understood not as discrete roles but as clusters of prototypical properties that tend to co-occur. Proto-Agent properties include volitional involvement in an event or state, sentience or perception, causal initiation of an event, and self-initiated movement or independent existence. Proto-Patient properties, by contrast, include undergoing a change of state, being causally affected by another participant, lacking control over the event, and being incrementally affected or consumed (Dowty, 1991).

Crucially, Proto-role theory adopts a gradient model of role assignment. Arguments are not assigned roles by satisfying a necessary-and-sufficient set of conditions; instead, role assignment emerges from a comparative evaluation of how many and which proto-properties an argument instantiates relative to other arguments of the same predicate. In Dowty's formulation, the argument with a greater concentration of Proto-Agent entailments is preferentially mapped to subject position, while the argument with more Proto-Patient entailments tends to be realized as an object or complement. However, this mapping is probabilistic rather than deterministic, allowing for variation across constructions and languages.

This entailment-based approach offers several theoretical advantages. First, it accounts naturally for argument alternations and borderline cases in which traditional thematic role labels are difficult to apply. Second, it provides a unified semantic explanation for cross-linguistic variation in argument realization, without positing language-specific role inventories. Third, it allows semantic role assignment to remain stable even when syntactic configurations change, as in passivization or argument reordering.

Proto-role theory is particularly well suited to analytic languages such as Khmer, where overt morphosyntactic cues to grammatical relations—such as case marking or agreement—are minimal or absent. In such languages, argument identification and role interpretation must rely heavily on lexical semantics, event structure, and contextual inference. By grounding role assignment in semantic entailments rather than surface syntax, Proto-role theory provides a principled framework for analyzing argument structure in contexts where syntactic diagnostics are weak or ambiguous.

Furthermore, the flexibility of the Proto-role framework makes it applicable to a wide range of constructions that pose challenges for traditional SRL approaches, including serial verb constructions, causatives, passives, and psychological predicates. In these constructions, arguments may exhibit mixed or non-canonical properties, such as abstract affectedness or metaphorical change. Proto-role theory accommodates such cases by allowing Proto-Agent and Proto-Patient properties to extend beyond strictly physical or eventive domains.

In the present study, Dowty's Proto-role theory serves as the primary theoretical foundation for semantic role analysis in Khmer. Proto-role entailments are operationalized as an explicit checklist of semantic properties, and role assignment is determined through a comparative evaluation of these properties across arguments. This approach enables a systematic and theoretically grounded analysis of Khmer argument structure, while remaining compatible with future extensions to computational semantic role labeling frameworks.

Analysis and Findings

Distribution of Proto-Role Entailments across the Dataset

Across the annotated dataset, a recurrent correlation can be observed between clusters of semantic entailments and Proto-role classification. Arguments exhibiting volitional control (VOL), sentience (SENT), and event initiation or causal involvement (CAUS) consistently tend to pattern as Proto-Agents, while arguments

characterized by affectedness (AFF), change of state (CHG), or lack of control (−CTRL) tend to pattern as Proto-Patients.

This generalization holds across both literary and journalistic sources, despite differences in genre, register, and discourse structure. Importantly, Proto-role assignment remains stable even in the absence of overt morphological case marking or agreement, suggesting that semantic entailments rather than surface syntax play a primary role in argument classification in Khmer.

Prototypical Agenthood in Intentional Action Predicates

In clauses describing intentional motion or interaction events, the subject argument systematically instantiates a dense cluster of Proto-Agent entailments. For example, in sentences such as “គេនឹងទៅជួបជុំ ឪពុកម្តាយ បងប្អូនញាតិសន្តាន”, the subject គេ exhibits [+VOL], [+SENT], [+MOVE], and [+CAUS], while the entities being met instantiate [+SENT] and [+AFF] without volitional control.

The absence of morphological marking does not impede role identification; instead, agency is inferred from the lexical semantics of the predicate complex (ទៅ + ជួប), which encodes intentional motion followed by social interaction. These cases align closely with Dowty’s prototypical Agent profile and represent the most stable Proto-Agent configurations in the dataset.

Abstract and Cognitive Proto-Patients in Learning and Mental Predicates

Several predicates in the dataset involve non-physical change, particularly in learning and cognition contexts. In “នាងបានទៅរៀនអក្សរនៅសាលាវត្តពោធិវិហារ”, the subject នាង satisfies all core Proto-Agent entailments, while the theme អក្សរ does not undergo physical alteration. Nevertheless, it is affected through epistemic change, namely acquisition and internalization.

These cases motivate an extension of Proto-Patient affectedness beyond material change to include cognitive transformation. Rather than undermining the Proto-role framework, such examples illustrate its flexibility: Proto-Patienthood in Khmer may be instantiated through abstract domains of change, particularly with predicates of learning, perception, and mental activity.

Serial Verb Constructions and Proto-role

Stability Serial verb constructions (SVCs) constitute a central testing ground for Proto-role theory in Khmer. In multi-verb sequences such as “គេនឹងបានដេកអស់ពេល... បរិភោគ... ដើរលេង...”, a single subject argument is shared across multiple predicates, each encoding a distinct sub-event.

The analysis shows that Proto-Agent entailments associated with the subject tend not to be redistributed or substantially altered across serial verbs. Instead, a unified Proto-Agent interpretation extends across the entire predicate complex. Notably, many SVCs lack a clearly affected participant, resulting in Proto-Agent-only configurations. This asymmetry may present difficulties under role inventories that presuppose obligatory Agent–Patient pairing, but is naturally accommodated by a Proto-role approach.

Passive Constructions and Role Persistence

Passive clauses in the dataset demonstrate that Proto-role assignment is independent of syntactic position. In constructions marked by ត្រូវបាន or agentive ដោយ, the promoted subject consistently instantiates Proto-Patient entailments such as [+AFF] and [+CHG], while the demoted agent retains [+VOL] and [+CAUS].

For instance, in “ការកើនឡើងនេះ ត្រូវបានគាំទ្រ ដោយ...”, the nominalized event ការកើនឡើង lacks volitional control but undergoes change due to external causation, while the ដោយ-phrase introduces an intentional Proto-Agent. These patterns provide preliminary support for separating semantic role assignment from grammatical subjecthood in Khmer.

Psychological Predicates and Metaphorical Affectedness

Psychological constructions present more gradient cases of Proto-role assignment. In “ខ្ញុំមានចិត្តមមែតិតមកបង”, the experiencer ខ្ញុំ retains Proto-Agent properties related to sentience and intentional mental activity, while បង functions as the target of emotional orientation.

Here, Proto-Patienthood is instantiated metaphorically through affective impact rather than physical change. While such assignments remain theoretically debatable, they are consistent with Dowty’s entailment-based model and reveal how Khmer encodes psychological relations through semantic, rather than morphosyntactic, cues.

Summary of Recurrent Patterns

Three recurrent patterns emerge from the analysis:

- Proto-Agent roles are strongly associated with volition, sentience, and control, regardless of syntactic position or construction type.
- Proto-Patient roles exhibit greater semantic variability, encompassing physical, cognitive, and metaphorical affectedness.
- Complex constructions—including SVCs, passives, and pro-drop structures—do not disrupt Proto-role entailments, suggesting relative semantic stability within the analyzed dataset across diverse syntactic environments.

Together, these findings demonstrate that entailment-based role assignment may offer a useful analytical framework for Khmer, where surface morphosyntax provides limited guidance for semantic interpretation.

DISCUSSION

The findings of this study engage directly with Dowty’s (1991) proposal that argument selection is shaped by clusters of semantic entailments rather than by fixed thematic labels. Within the limited dataset examined here, recurrent Proto-role tendencies can be observed across a range of Khmer constructions, particularly in contexts where overt morphosyntactic diagnostics remain weak or ambiguous.

From the perspective of semantic role labeling, the analysis highlights potential limitations of heavily syntax-dependent or predicate-specific annotation schemes when applied to analytic languages such as Khmer. While frameworks such as PropBank and FrameNet have proven highly successful for well-resourced languages, their implementation typically presupposes extensive annotated corpora, stable parsing infrastructure, or large predicate inventories. Proto-role entailments, by contrast, may provide a lighter-weight semantic layer that can support early-stage annotation efforts in under-resourced settings.

The Khmer data also suggest that Proto-role analysis may remain relatively stable across constructions involving serial verbs, passivization, and argument omission. In particular, the persistence of entailment patterns across passivization supports approaches that distinguish semantic role assignment from grammatical subjecthood. Similarly, the treatment of serial verb constructions as unified predicate complexes appears compatible with an entailment-based interpretation of argument structure.

At the same time, the present findings should be interpreted cautiously. The dataset remains small, genre coverage is limited, and annotation was conducted by a single researcher. The study therefore does not claim to provide a comprehensive account of Khmer semantic roles or a validated computational SRL framework. Instead, it should be understood as a feasibility-oriented pilot investigation intended to explore whether Proto-role analysis can serve as a productive foundation for future Khmer semantic annotation research.

More broadly, this work positions Proto-role analysis as a possible conceptual bridge between theoretical linguistics and future Khmer NLP development. Because Proto-role entailments can be operationalized without requiring extensive morphosyntactic marking, they may offer a practical starting point for semantic annotation in low-resource analytic languages where large-scale SRL infrastructure is not yet available.

CONCLUSION

This study has explored the applicability of Dowty's Proto-role theory to Khmer argument structure through a small-scale manually annotated dataset. The analysis suggests that Proto-Agent and Proto-Patient entailments may provide a useful semantic framework for examining argument realization in a morphologically impoverished language where grammatical relations are often weakly encoded morphosyntactically.

Across the analyzed examples, recurrent entailment tendencies were observed in constructions involving serial verbs, passives, psychological predicates, and argument omission. Although the findings remain preliminary, they indicate that Proto-role analysis may help capture aspects of Khmer semantic organization that are not easily represented through purely syntactic approaches.

Importantly, the present study should be understood as an exploratory feasibility study rather than a comprehensive SRL implementation. The limited dataset size, restricted genre coverage, and absence of inter-annotator validation constrain the generalizability of the findings. Future research should therefore expand the corpus to include conversational Khmer, spoken interaction, and digital communication data, while also incorporating multi-annotator evaluation procedures and comparative analysis with alternative SRL frameworks such as PropBank and FrameNet.

While no computational model was developed in the present study, the analysis outlines several conceptual directions that may support future research in Khmer semantic annotation, dependency-based SRL, machine translation, and syntactic–semantic parsing. More broadly, the study aims to contribute to ongoing discussions concerning how semantic role frameworks can be adapted to low-resource analytic languages whose grammatical organization differs substantially from that of well-studied Indo-European languages.

REFERENCES

1. Aikhenvald, A. Y., & Dixon, R. M. W. (Eds.). (2006). *Serial verb constructions: A cross-linguistic typology*. Oxford University Press.
2. Baker, M. C. (2003). *Lexical categories: Verbs, nouns and adjectives*. Cambridge University Press.
3. Carnie, A. (2013). *Syntax: A generative introduction* (3rd ed.). Wiley-Blackwell.
4. Dowty, D. R. (1991). Thematic proto-roles and argument selection. *Language*, 67(3), 547–619. <https://doi.org/10.2307/415037>
5. Fillmore, C. J., Johnson, C. R., & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3), 235–250. <https://doi.org/10.1093/ijl/16.3.235>
6. Haiman, J. (2011). Khmer. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. <https://wals.info>
7. He, L., Lee, K., Lewis, M., & Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 473–483). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1044>
8. Huynh, S. (2023). *Serial verb constructions in Khmer* (Unpublished doctoral dissertation).
9. Johansson, R., & Nugues, P. (2008). Dependency-based semantic role labeling of PropBank. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 69–78). Association for Computational Linguistics.
10. Kratzer, A. (1996). Severing the external argument from its verb. In J. Rooryck & L. Zaring (Eds.), *Phrase structure and the lexicon* (pp. 109–137). Springer. https://doi.org/10.1007/978-94-015-8617-7_5
11. Levin, B., & Rappaport Hovav, M. (2005). *Argument realization*. Cambridge University Press.
12. Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71–106. <https://doi.org/10.1162/0891201053630264>

13. Pearson, M. (2013). The syntax and semantics of argument structure in Khmer (Doctoral dissertation, University of California, Los Angeles). ProQuest Dissertations Publishing.
14. Self, S. (2004). Serial verb constructions in Mon-Khmer languages. *Mon-Khmer Studies*, 34, 1–22.
15. Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (2007). Combination strategies for semantic role labeling. In *Proceedings of the Human Language Technology Conference of the NAACL* (pp. 105–112). Association for Computational Linguistics.
16. Thu, Y. K., Ding, C., & Utiyama, M. (2016). A parallel corpus for Asian languages: ALT project. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1626–1632). European Language Resources Association.
17. Williams, A. (2015). *Arguments in syntax and semantics*. Cambridge University Press.

APPENDIX

To operationalize Proto-role assignment, we employed an entailment checklist adapted from Dowty (1991), summarized in Table 1. The checklist was applied uniformly across simple predicates and predicate complexes in serial verb constructions.

Table 1. Proto-role entailment checklist used for Khmer annotation (after Dowty, 1991)

Code	Entailment feature	Operational definition (annotation criterion)	Proto-role tendency
VOL	Volitionality	Argument intentionally initiates or controls the event or mental activity	Proto-Agent
SENT	Sentience	Argument is capable of perception, cognition, or emotional experience	Proto-Agent
CAUS	Causation	Argument brings about an event or change in another participant	Proto-Agent
MOVE	Self-initiated motion	Argument undergoes or controls motion independent of external force	Proto-Agent
CTRL	Control	Argument exercises control over the unfolding of the event	Proto-Agent
AFF	Affectedness	Argument undergoes impact, influence, or alteration due to the event	Proto-Patient
CHG	Change of state	Argument experiences a physical, cognitive, or abstract state change	Proto-Patient
¬CTRL	Lack of control	Argument does not control the event affecting it	Proto-Patient
INC	Incremental theme	Argument is gradually affected or consumed over the course of the event	Proto-Patient
PSY-AFF	Psychological affectedness	Argument is affected emotionally, cognitively, or attentively (non-physical)	Proto-Patient

Role Assignment Principle (used Throughout Section 6)

Proto-role classification is determined by **comparative dominance of entailment features** rather than by any single diagnostic.

An argument is classified as **Proto-Agent** if it instantiates a denser or more salient cluster of Proto-Agent-oriented entailments (VOL, SENT, CAUS, MOVE, CTRL).

An argument is classified as **Proto-Patient** if it instantiates Proto-Patient-oriented entailments (AFF, CHG, ¬CTRL, INC, PSY-AFF), even in the absence of physical change.