A Scalable Tool for Measuring Manner and Result Verbs in Developmental Language Research

Divyesh Pratap Singh
University at Buffalo

Dakshesh Gusain
Nanyang Technological University

Federica Bulgarelli
University at Buffalo

Alison Eisel Hendricks
University at Buffalo

John Beavers
The University of Texas at Austin

Nathan M. Beers
University at Buffalo

Ifeoma Nwogu
University at Buffalo
inwogu@buffalo.edu

Abstract

Manner and result verbs encode different aspects of event structure and have been discussed in developmental work as a potentially informative distinction for studying early verb learning. However, this distinction remains difficult to measure at scale because large annotated resources for manner and result classification are not currently available. We present a computational approach for identifying manner and result verbs in sentence context. Using linguistically informed prompts, we generate sentence-level annotations with large language models over data drawn from MASC and InterCorp, extending coverage from previously annotated portions of VerbNet to 436 classes. We then train a RoBERTa-based classifier on these annotations and evaluate it on three held-out gold-standard datasets, including previously annotated items and a new expert-annotated set. Across these evaluations, the model shows promising performance, with average accuracy up to 89.6%. We present this work as a scalable measurement tool that can support future research on verb semantics in developmental and other language datasets, while noting that further validation is needed for borderline cases, mixed manner/result verbs, and downstream developmental applications.

1 Introduction

Early language development depends not only on how much language children hear, but also on the kinds of meanings encoded in the words they learn. Verbs are especially important because they support children's transition to multiword speech and later grammatical development. Verb vocabulary around age two predicts later grammatical outcomes and, for some developmental questions, may be more informative than noun vocabulary (Hadley et al., 2016). These issues are especially relevant for late talkers, whose early language trajectories are heterogeneous and whose later outcomes are difficult to predict from broad lexical measures alone.

One semantic distinction that has become relevant in this literature is the contrast between *manner* and *result* verbs. Manner verbs encode how an action is carried out (e.g., *rub*, *scribble*, *run*), whereas result verbs encode an outcome or change of state (e.g., *clean*, *fill*, *open*) (Hovav and Levin, 2010; Levin, 2008). Developmental work suggests that this distinction may be informative for understanding variation in early verb learning. For example, Horvath et al. (2022) report that the relative proportions of manner and result verbs differ between late talkers and typically developing children, and that children who produce more manner verbs also tend to produce more verbs overall (Horvath et al., 2019, 2022). Because a substantial proportion of children with early language delay later meet the criteria for Developmental Language Disorder (DLD), the task of identifying finer-grained semantic properties of children's early vocabularies may help clarify which aspects of early language are associated with these later outcomes.

At the same time, this distinction remains difficult to study at scale. Although computational linguistics has made substantial progress on grammatical annotation tasks such as part-of-speech tagging (DeRose, 1988), fine-grained semantic categorization is generally more challenging. Prior work on related event-semantic distinctions suggests that verb meaning is difficult to classify automatically in context (Friedrich et al., 2022; Friedrich and Gateva, 2017; Metheniti et al., 2022; Friedrich et al., 2016). As a result, theoretically important contrasts such as manner versus result verbs still lack broad, scalable annotation resources, limiting their use in developmental language research.

To address this gap, we present a computational approach for identifying manner and result verbs in context. We use large language models (LLMs) as informed annotators, drawing on established linguistic definitions of manner and result verbs together with a small set of illustrative examples. We

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 117–128
July 3, 2026

prompt LLMs to label sentences from the Manually Annotated Sub-Corpus (MASC; Ide et al. 2008) and the InterCorp parallel corpus (ek Čermák and Rosen 2012), expanding coverage from 151 previously annotated VerbNet classes to 436 classes (Brown et al., 2019; Kipper et al., 2008). We then fine-tune a pretrained RoBERTa classifier (Liu et al., 2019) on these labels and evaluate it on three held-out gold-standard datasets. We position this system as a scalable measurement tool that can support research on verb semantics in larger corpora, including developmental language data.
In summary, our contributions are:

- We present a scalable computational framework for identifying *manner* and *result* verbs in sentence context, enabling this theoretically important distinction to be measured in larger language datasets.
- We introduce an annotation pipeline that leverages large language models to generate training data for this task in the absence of large-scale gold-standard resources.
- We show that a RoBERTa-based classifier trained on these annotations can reliably distinguish manner and result verbs in context.
- We will publicly release our code and annotated dataset, extending coverage to 436 VerbNet classes, to support future research.

2 Understanding Verb Root Meaning

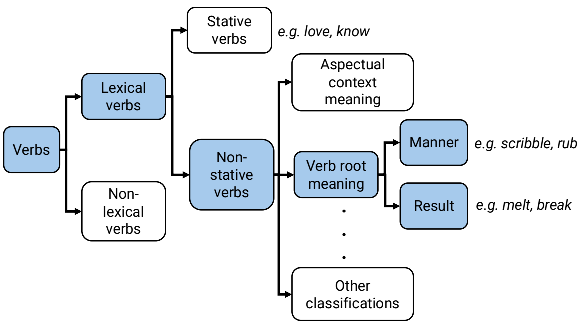

Figure 1: Hierarchy of verb classification, with manner and result verbs as subdivisions of non-stative verbs

Figure 1 shows the hierarchy of verb classifications relevant to our proposed task. At a high level, lexical verbs can be categorized into **stative** and **non-stative** verbs.

- **Stative verbs** describe a continuous or unchanging state rather than an action or event, e.g. *love* in the sentence "She loves her dog,"
- **Non-stative verbs**, on the other hand, describe actions or events that unfold over time and can lead to changes in state.

Non-stative verbs can be further classified based on different linguistic properties, such as **aspectual features** (e.g., telicity, durativity) and **argument realization patterns** (e.g., causative-inchoative alternation),etc. However, a fundamental classification based on the **inherent meaning stored in the verb root** is the difference between manner verbs and result verbs (Levin and Hovav, 1991; Hovav and Levin, 2010; Levin, 2008); This distinction plays a significant role in both language acquisition (Gentner and Boroditsky, 2001) and the way verbs encode event semantics.

- **Manner verbs** specify *how* an action is performed but do not encode its outcome (e.g., *scribble, rub, sweep, flutter*).
- **Result verbs** specify *what* change or outcome occurs, without specifying how the action is carried out (e.g., *clean, melt, fill, arrive*).

Unlike classifications such as telicity, which are determined at the clause level (Friedrich and Gateva, 2017), the manner/result distinction is typically analyzed as a property of the verb root (Levin, 2008), meaning that it is expected to remain relatively stable across contexts.

2.1 Illustrating the difference between manner and result verbs

To understand this complementarity, consider the following pair of sentences:

1. *Anna shoveled the snow.*
2. *Anna cleared the snow.*

In (1), the verb *shoveled* focuses on *how* the action was performed, i.e. the process of moving the snow with a shovel, but does not guarantee that the snow was removed. In contrast, in (2), the verb *cleared* encodes the outcome, that the snow was removed, but does not specify how Anna accomplished this (she could have used a shovel, a snowblower, or even melted it). This distinction is crucial because it shows that result verbs inherently encode a outcome, while manner verbs focus on the process. One way to test whether a verb encodes a result or manner is by using the *denying the result* diagnostic test (Hovav and Levin, 2010). If the sentence remains logical, the verb does not inherently encode a result:

Anna shoveled the snow, but the snow is still there. (logical)

Since this sentence makes sense, we can infer that "*shovel*" does not encode a result; it only describes the action. Thus even though real-world knowledge might suggest that performing an action in a certain way will typically lead to a result, this is not always true. The **core meaning of a verb remains stable across different contexts**. However, trying the same test with a result verb leads to contradiction:

Anna cleared the snow, but the snow is still there. (contradiction)

3 Manner and Result Verb Diagnostics

To effectively transfer the knowledge of result and manner heuristics into an LLM annotator, it is essential to identify the linguistic features that reliably distinguish them. Since the manner/result distinction is inherent to the verb root rather than being compositionally determined, much of this semantic information is encoded within the verb itself. However, sentence structure also offers useful cues, as manner and result verbs occur in complementary syntactic environments. In particular:

- Manner verbs frequently occur without a direct object.
- Result verbs typically require an object to specify the entity undergoing change.
- Only result verbs consistently participate in causative/inchoative alternations.

Below, we present these sentence formation diagnostics that linguistic researchers have leveraged for result and manner verb identification.

3.1 Sentence formation diagnostics

Diagnostic 1: Object omission Manner verbs can appear without a direct object, whereas result verbs typically require one (Hovav and Levin, 2010). Consider the following examples:

- Manner verb: *Anna wept all day.* (Acceptable without an object)
- Result verb: *The child broke _ ?* (Unacceptable without an object)

This suggests that manner verbs describe an action that can occur independently, whereas result verbs typically requiring an affected entity.

Diagnostic 2: causative/inchoative alternation The causative/inchoative alternation refers to a pattern in which a verb appears both in a causative form (with an explicit agent) and an inchoative form (where the event occurs spontaneously without an agent)(Hovav and Levin, 2010; Beavers and Koontz-Garboden, 2012; Levin and Hovav, 1991). This alternation serves as a reliable test for result verbs, as manner verbs rarely allow such transformations.

- Result Verb:
  - *Causative: The child broke the vase.* (An agent explicitly causes the event.)
  - *Inchoative: The vase broke.* (The event occurs without an explicit agent.)
- Manner Verb:
  - *Causative (transitive): John wiped the table.*
  - *Inchoative (intransitive): The table wiped.* (Ungrammatical)

Unlike result verbs, manner verbs describe a process but do not inherently encode an endpoint. As a result, they resist appearing in inchoative constructions.

3.2 Semantic Diagnostics: beyond syntactic patterns

While the above syntactic tests provide useful heuristics, they are not always sufficient for classification. Certain verbs such as *climb*, and *cut* resist strict categorization due to polysemy or context-dependent interpretations (Levin, 2008; Beavers and Koontz-Garboden, 2012). To address this, researchers have therefore investigated **semantic properties** that further refine the manner/result distinction.

Diagnostic 3: Telicity Telicity refers to whether a verb's action has a natural endpoint or goal. A verb is *telic* if it describes an action that reaches completion, such as *build* or *paint* (*She built a house.*, *He painted a portrait.*). These actions have a defined conclusion. In contrast, a verb is *atelic* when the action is ongoing, lacks a specific endpoint, or its completion is uncertain, as seen with verbs like *know* or *sleep* (*She knows the answer*, *They slept peacefully*). Dowty (2012); Levin and Hovav (1991); Krifka (1992) observed a correlation between result verbs and telicity. However, while result verbs involving two-point changes (e.g., arrive, reach, die, crack, find) are necessarily telic, result verbs describing degree achievements verbs (cooled, heat) are not strictly telic. Consider the shift in telicity with a time modifier.

- *The dryer dried the clothes for two hours* (Atelic: no clear endpoint)
- *The dryer dried the clothes in two hours* (Telic: the drying is completed)

Diagnostic 4: scalar vs. non-scalar changes Hovav and Levin (2010) proposed that the distinction between **scalar** and **non-scalar** changes provides a strong basis for differentiating manner and result verbs. Since both verb types denote dynamic events, they inherently involve a change (Dowty, 2012); however, the nature of that change differs. Result verbs are characterized by changes that occur along a measurable scale, either as a two-point change (e.g., break) or as a gradable change (e.g., melt). In contrast, manner verbs involve non-scalar changes that cannot be readily quantified along a single dimension. For example, the action described by the verb *flap* entails a complex, multidimensional movement that is not easily measurable. Result verbs thus describe changes along a measurable scale, meaning the event involves a progression toward a defined endpoint.

- Two-point scale (binary change): *break, die, arrive*
- Gradable scale (continuous change): *melt, cool, widen*

Manner verbs describe non-scalar changes, where the event unfolds without a well-defined trajectory.

- Example: *flap, jog, scribble*-these actions involve repeated or multidimensional motion rather than progression toward an endpoint.

This distinction supports Levin (2008) hypothesis of manner-result complementarity, which posits that a single action cannot simultaneously encode both a scalar and non-scalar change.

3.3 Implications for LLM annotation

The manner/result distinction is semantically encoded (as part of the verb meaning), but syntactic diagnostics contribute in testing participation of a verb in a particular category using sentence structure. These distinctions are integrated into our approach by structuring our prompt designs around the sentence formation rules (diagnostic tests 1 and 2) and semantic features (diagnostic tests 3 and 4).

4 Methodology

In this section, we describe the end-to-end pipeline, including (i) annotation of training data for manner/result labels, (ii) training of the tagging model, and (iii) construction of held-out (gold-standard) evaluation datasets.

As explained in the Contributions list from Section 1, to the best of our knowledge, this is the first attempt to computationally annotate and classify texts using the manner/result constructs. For this reason, there are no known annotated datasets useful for training a computational model. Hence, to address this challenge, we resorted to LLMs, to assist in creating a large, annotated dataset with result and manner verb labels. He et al. (2023); Zhang et al. (2023) showed that with structured prompts and few-shot examples, LLMs can effectively mimic human annotations for various NLP tasks.

4.1 LLM-Based training data annotation

For this task, we compile the sentences from MASC and InterCorp dataset consisting of 4,492 sentences and 2,554 unique verb occurrences. Next, using our expert-guided prompts, we use the GPT-4o model to identify the non-stative verbs in each sentence and classify them based on our manner-result diagnostic framework. The rules for designing the two separate prompts for GPT-4o, where each focuses on a different aspect of verb classification, are described:

Prompt 1 (semantic properties): checks for scalar vs. non-scalar change information embedded within verb root. The two major rules driving Prompt 1 are shown in Figures 3 and 4.

Prompt 2 (sentence structure): emphasizes possible sentence formation patterns, including object omission and causative/inchoative alternations. Due to space constraints, the governing rules for Prompt 2 is presented in the Appendix B.

The prompts provided to LLM yielded 4,928 result verbs, 4,247 manner verbs, and 64,767 other words tagged with other categories such as nouns, determiners, pronouns, etc.

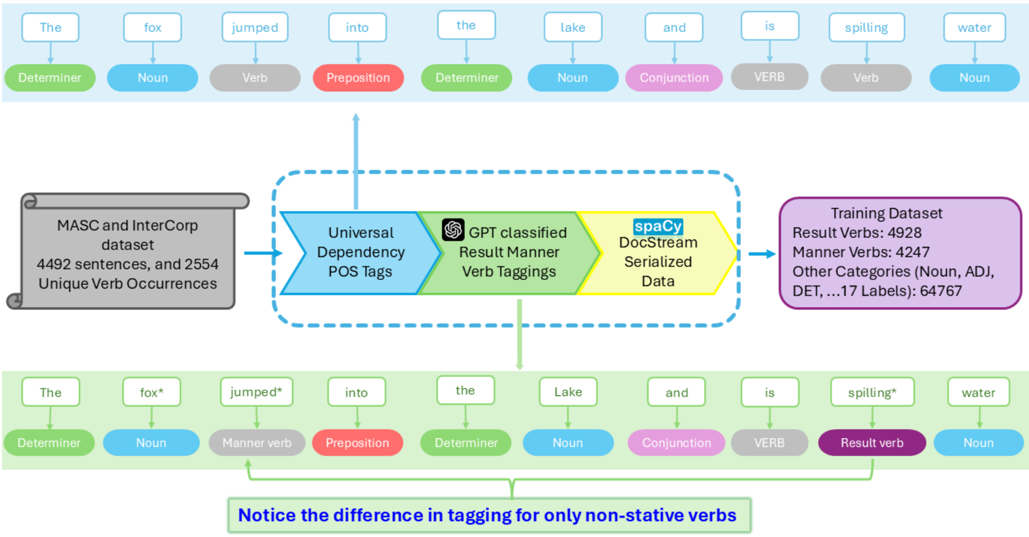

Figure 2: Overview of our data generation pipeline.

Manner verbs (manner):
Definition: These verbs encode the *how* of an action, focusing on the method or pattern without specifying an outcome.
Semantic Basis: These verb often involve **nonscalar** or **complex** actions that are often multidimensional (e.g., the specific pattern of leg movements while jogging which is complex and a culmination of multiple actions.)
Usage: "She **ran** towards the market." (Focus on how she went to the market)

Result verbs (result):
Definition: These verbs encode the *outcome* or resultant state that follows from an action.
Semantic Basis: Involve **scalar** changes that occur along a defined scale (e.g., temperature increasing, distance decreasing).
Usage: "He **melted** the ice." (Focus on the fact that **melting** alone stores all the information that something has changed form)

Figure 3: Result Manner Verb Definition

Verb Root Classification
Definition: The verb has to be classified based on its primary lexical meaning and the inherent information that the verb independently encodes irrespective of the context.

Example: *"Wipe"*
Sentence: "He **wiped** the table **clean**."
The verb *wipe* primarily indicates the manner of cleaning; the resulting state (*"clean"*) is introduced by the adjective *clean*, not by *wipe* itself. Therefore, the verb *wipe* remains a Manner Verb because the outcome (i.e., making something clean) is not inherently encoded by the verb's own meaning.

Past Information in Verb Classification:
Manner Verbs inherently encode the way an action was performed in past instances, while Result Verbs do not.

Example 1: ***Cook*** *(Result Verb)*
Sentence: "She **cooked** chicken for him."
Rewriting it as: "She **cooked** chicken for him again." The word **cook** does not provide information about how the food was cooked before, it can be grilled, sautéed, etc.

Example 2: ***Sauté*** *(Manner Verb)*
Sentence: "She **sautéed** chicken for him."
Rewriting it as: "She **sautéed** chicken for him again." The verb **sauté** stores the information that the chicken was previously also cooked using sautéeing.

Figure 4: Verb Root Classification

4.2 Approaching the problem as part-of-speech (POS) tagging

Since our task involves both verb classification and detection them in a sentence, we adopt a sequence-tagging approach, similar to part-of-speech (POS) tagging, rather than formulating it as a binary classification task. This enables us to identify non-stative verbs, since modal and auxiliary verbs are readily identifiable using syntactic structures.

The advantages of taking a sequence-tagging approach include:

1. Explicit identification of non-stative verbs: By tagging all the words in a sentence, we can reduce the final error by isolating and classifying only the non-stative verbs, thus avoiding any misclassification of auxiliary and modal verbs (e.g., *can, might, have, be*).

2. Facilitates our ultimate goal in child language research applications: Our model can be directly integrated into the child language research pipeline where most often the goal is to scan through the complete sentences spoken by a child, and identify the number of result and manner verbs. Tagging only the non-stative verbs eliminates an additional step to filter any stative and non-lexical verbs.

Figure 2 illustrates the sequence-tagging based data generation pipeline. First, we tag each sentence using any standard POS tagger. For example, the sentence "*The fox jumps into the lake and is spilling water*" is initially tagged as:
"*The* (DT) *fox* (NN) *jumps* (VB) *into* (IN) *the* (DT)

lake (NN) *and* (CC) *is* (VB) *spilling* (VB) *water* (NN)."

Next, we update the tagging for non-stative verbs using GPT (Achiam et al., 2023), classifying them as either result or manner verbs. The modified tagged dictionary:

"*The* (DT) *fox* (NN) *jumps* (manner) *into* (IN) *the* (DT) *lake* (NN) *and* (CC) *is* (VB) *spilling* (result) *water* (NN)."

This process is applied to all sentences, and finally compiled to create the training set.

4.3 Curation of gold-standard test data

We evaluate our models on three held-out gold-standard datasets. The first, the *Linguists verb-root* dataset, contains 83 verbs (34 result, 49 manner) compiled from prior work on lexical semantics and verb-root classification (Levin, 2008; Hovav and Levin, 2010; Beavers and Koontz-Garboden, 2012; Levin and Hovav, 1991). The second, the *Psycholinguistic verb-root* dataset, comes from Horvath et al. (2022), who annotated 77 MacArthur-Bates CDI verbs (36 result, 41 manner).

Because these datasets together covered only 151 of 487 VerbNet classes, we created a third set with the help of an expert linguist. Guided by VerbNet, we constructed 200 new sentences spanning 346 classes; the expert labeled 48 as result, 62 as manner, 23 as stative, and 67 as unsure. We refer to this as the *Expert-annotated verb-root* dataset (see Appendix A).

All 3 datasets are distinct from the MASC training data and are used only for held-out evaluation.

5 Computational Modeling

This section outlines our computational approach for classifying verbs according to both *manner/result* and *stative/non-stative* properties.

5.1 Model architecture

Our tagging pipeline is implemented using a spaCy wrapper (Honnibal and Montani, 2017) and follows a sequence of components as shown in Figure 5: (1) a tokenizer, (2) fine tuning a pre-trained transformer-based feature extractor, (3) a feature selector (pooling layer), and (4) a classification head.

Tokenizer. Byte Pair Encoding Tokenization (Sennrich, 2015) strategy segments raw text into tokens, and matches with our downstream RoBERTa model default tokenization strategy.

Transformer. We employ **RoBERTa-base** model (125 million parameters) as the backbone of our pipeline, which encodes each token - in conjunction with its context - into a contextualized representation.

Feature Selector. To reduce subword embeddings to a single vector per token, we apply mean pooling (`reduce_mean.v1`).

Classifier. We use label smoothing (0.05) to predict token-level labels for default parts-of-speech tagging (17) plus two new labels (result and manner) Each token's pooled embedding is projected into logits corresponding to these classes, and the model is optimized via cross-entropy loss.

5.2 Feature representation

Contextual embeddings. Tokens are generated using BPE tokenizer that sequences via pretrained *RoBERTa-base* vocabulary. This aids in capturing syntactic signals.

Token-Level pooling. Mean pooling operation over subword embeddings yields 768-dimensional vectors representing token-level features. A feature selector (`TransformerListener`) is applied to remove redundant information, reducing them to 300-dimensional representations, retaining semantic and syntatic features.

5.3 Training procedure

Hyperparameters. We train the model using Adam with learning rate $= 5 \times 10^{-5}$, $\beta_1 = 0.9, \beta_2 = 0.999$, weight decay (L2) $= 0.01$, gradient clipping $= 1.0$ and batch size $= 128$.

The model is trained for up to 20,000 steps, with evaluation every 200 steps. A patience of 1,600 steps is used to halt training if the validation accuracy fails to improve. This setup balances thorough exploration of the parameter space with computational efficiency.

All experiments run using a word-based batcher and compounding batch sizes (start=100, stop=1000, compound=1.001) on a single GPU (NVIDIA RTX A6000) for 25 minutes training time. The final checkpoint is selected based on the highest tagging accuracy on our gold annotated dataset.

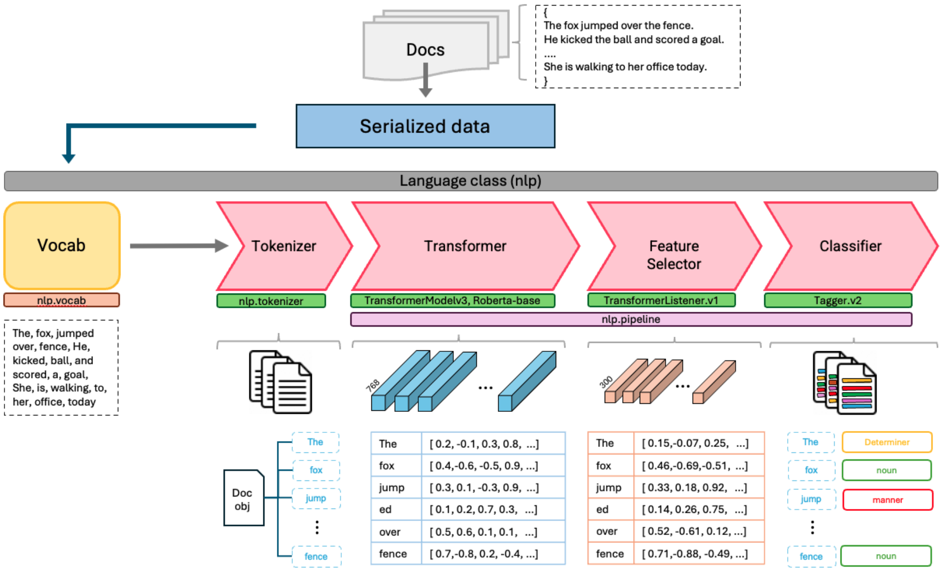

Figure 5: Overview of model architecture

	Acc.	F_1 (result)	Precision (result)	Recall (result)	F_1 (manner)	Precision (manner)	Recall (manner)
Model 1 (Trained using Prompt 1)							
Linguistic dataset	0.94	0.93	0.89	0.97	0.95	0.98	0.92
Psycholinguistic dataset	0.90	0.88	1.00	0.78	0.91	0.84	1.00
Expert-annotated dataset	0.86	0.85	0.84	0.85	0.88	0.89	0.87
Model 2 (Trained using Prompt 2)							
Linguistic dataset	0.94	0.93	0.91	0.94	0.95	0.96	0.94
Psycholinguistic dataset	0.84	0.80	1.00	0.67	0.87	0.77	1.00
Expert-annotated dataset	0.81	0.80	0.82	0.77	0.84	0.84	0.84

Table 1: Comparison of Model 1 and Model 2 on different datasets.

6 Experiments and Results

We evaluate our models on the three gold-standard datasets described in Section 4.3, the Linguist, Psycholinguists and Expert-annotated verb root datasets.

Quantitative Results We trained our model using annotations generated from two distinct prompts -one emphasizing the semantic properties of verbs and the other focusing on sentence structure. Table 1 presents model performance across multiple datasets, highlighting accuracy, F1-score, precision, and recall for result and manner verbs.

- Model 1 consistently outperforms Model 2 achieving equal or higher accuracy across all three datasets.
- The Linguistics dataset performed the best among all three test datasets and across the two prompts. This is likely due to the fact that we constructed our governing prompt rules based on information gleaned from the papers from which that dataset was culled.
- Model 1 shows weaker recall (0.67) for result verbs on the Psycholinguistic dataset, indicating higher misclassification rates. Inspection of the disagreements suggests that some verbs in this dataset (e.g., *paint*, *dump*, *drink*) may be borderline cases, for which psycholinguistic annotations and our linguistically guided framework assign different labels. Since the dataset in Horvath et al. (2022) was annotated for developmental research purposes, these mismatches may reflect differences in annota-

tion criteria across domains rather than simply model failure.Horvath et al. (2019) indicated in their paper that the authors annotated the verbs themselves.

The fact that Model 1 performs better than Model 2 suggests that understanding the semantic information inherent in verb roots is more crucial than analyzing sentence structure, for this verb categorization task.

7 Developmental Use Case

We illustrate the utility of our approach through its application to a longitudinal developmental dataset of parent-child interactions involving typically developing (TD) toddlers and Late Talkers (LTs). In this use case, transcripts from the CHILDES Clinical English Ellis Weismer corpus (Weismer et al., 2013) are processed with our classifier to identify manner and result verbs in caregiver and child speech at 30 months, yielding speaker-level measures such as verb types, tokens, and manner-to-result ratios. These measures can then be related to later language outcomes at 42 and 66 months, including MLU, TTR, and IPSyn scores. IPSyn was measured following Scarborough (1990), TTR was included as a standard index of lexical diversity (Hess et al., 1986), and mean length of utterance (MLU) was derived from examiner-child language samples at two time points to index later grammatical development (Weismer et al., 2013).

This use case is motivated by prior developmental findings suggesting that manner and result verbs may differ in their relation to language growth. For example, Horvath et al. (2022) report that TD children's vocabularies contain relatively more manner verbs, whereas Late Talkers' vocabularies contain relatively more result verbs; children who produce more manner verbs also tend to produce more verbs overall. At the same time, broad measures of parental input have not consistently distinguished the language environments of TD children and Late Talkers, suggesting that finer-grained semantic properties of the input may also be informative (D'Odorico and Jacob, 2006; Naigles and Hoff-Ginsberg, 1998).

We view this as an example of the kind of developmental analysis that scalable manner/result classification supports. Rather than relying only on coarse measures of input quantity, researchers can use the present tool to examine whether the semantic composition of caregiver and child verb use is related to later language outcomes. In this sense, the classifier functions as a corpus-based measurement tool that may help support richer analyses of early verb learning and developmental variation.

This use case is intended as an illustration of research utility rather than a clinical application. Although the tool supports corpus-based measurement of a theoretically motivated semantic distinction, further validation will be needed before drawing stronger conclusions about diagnostic use or generalization across developmental datasets (Verhage et al., 2020; Conti-Ramsden et al., 2018).

8 Conclusion

We present a computational approach to identifying manner and result verbs in context. By using LLM-generated annotations, we expand coverage from 151 to 436 VerbNet classes and train a RoBERTa-based classifier on this distinction.

The model achieves up to 89.6% average accuracy across three gold-standard evaluation sets (with annotations by expert linguists). Our results suggest that semantic properties of non-stative verb roots contribute more to this task than sentence structure alone, supporting the value of linguistically informed modeling for event-structure classification.

Future work will test the approach on more diverse data, extend it to other languages, and further explore its use in developmental language research. At present, we see the model as a tool for corpus-based analysis which can have developmental relevance, rather than as a direct clinical or diagnostic decision-making tool.

Limitations

The following section illustrates some of the current limitations of the proposed research:

- In this work, although we have identified comprehensive sets of manner/results verb diagnostics, and have used these to construct intelligent prompt for generating our training data, *we did not consider polysemous verbs and subtle alternations of verbs*.

- While LLMs perform well in verb categorization, they rely on statistical associations rather than linguistic principles, and this could lead to inconsistencies. *When a random sampling of the resulting annotated data was "spot-checked" by an expert, the LLM annotations were not 100% accurate*.

- Subsequent analyses by Beavers and Koontz-Garboden (2012) noted that certain verbs exhibit both manner and result properties. For instance, the verb *guillotine*, and *drowned* explicitly convey the manner of killing (i.e., how the action is performed) while also implying the resultant state (i.e., that the person is killed). Similar behavior is observed with certain cooking verbs such as *braise*, *sauté*, and *poach*. However, *for our analysis in this work, we focused only on the manner or result aspect of non-stative verbs*.

- A critical challenge in this work was the scarcity of expertise in the research area, with only a handful of specialists available. We therefore relied mainly on one expert to create our gold-standard expert annotation and *we were unable to obtain inter-rater reliability*.

Ethical Considerations

This work raises several ethical considerations relevant to computational tools for semantic annotation.

- Large language models may reproduce linguistic and cultural biases present in their training data. Consequently, our annotation pipeline may be less accurate for speakers whose language use is underrepresented in standard corpora, including speakers of regional or non-standard varieties of English.

- We view the present system as a research tool rather than a clinical or diagnostic instrument. Any future use in developmental or clinical contexts would require substantial additional validation across populations, settings, and language varieties.

- Since the current data are drawn primarily from standard English sources, the model may not generalize equally well to all communities. Expanding evaluation to more diverse dialects and speech contexts is therefore an important direction for future work.

Broader Impacts

This work has potential broader impacts across developmental language research, linguistics, and computational modeling.

- For developmental research, scalable measurement of manner and result verbs may support more fine-grained analyses of early language development, including studies of late talkers and children at risk for persistent language difficulties. Because later outcomes among children with early language delay are heterogeneous, tools that make theoretically motivated semantic distinctions measurable in larger corpora may help researchers better characterize variation in children's early vocabularies and language input (Catts et al., 2012; Conti-Ramsden et al., 2018).

- For computational linguistics, this work provides an example of how linguistic theory and domain expertise can be incorporated into annotation pipelines and downstream modeling. In particular, the use of linguistically informed prompts illustrates one possible strategy for generating training data in tasks where large gold-standard semantic resources are not yet available.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

John Beavers and Andrew Koontz-Garboden. 2012. Manner and result in the roots of verbal meaning. *Linguistic inquiry*, 43(3):331–369.

Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2019. Verbnet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163.

Hugh W Catts, Donald Compton, J Bruce Tomblin, and Mindy Sittner Bridges. 2012. Prevalence and nature of late-emerging poor readers. *Journal of educational psychology*, 104(1):166.

Gina Conti-Ramsden, Kevin Durkin, Umar Toseeb, Nicola Botting, and Andrew Pickles. 2018. Education and employment outcomes of young adults with a history of developmental language disorder. *International journal of language & communication disorders*, 53(2):237–255.

Steven J. DeRose. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.

Laura D'Odorico and Valentina Jacob. 2006. Prosodic and lexical aspects of maternal linguistic input to late-talking toddlers. *International Journal of Language & Communication Disorders*, 41(3):293–311.

David R Dowty. 2012. *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*, volume 7. Springer Science & Business Media.

Franti ek Čermák and Alexandr Rosen. 2012. The case of intercorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2022. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. *arXiv preprint arXiv:2208.09012*.

Dedre Gentner and Lera Boroditsky. 2001. Individuation, relativity, and early word learning. *Language acquisition and conceptual development*, 3:215–256.

Pamela A Hadley, Matthew Rispoli, and Ning Hsu. 2016. Toddlers' verb lexicon diversity and grammatical outcomes. *Language, speech, and hearing services in schools*, 47(1):44–58.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Carla W Hess, Karen M Sefton, and Richard G Landry. 1986. Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research*, 29(1):129–134.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Sabrina Horvath, Justin B Kueser, Jaelyn Kelly, and Arielle Borovsky. 2022. Difference or delay? syntax, semantics, and verb vocabulary development in typically developing and late-talking toddlers. *Language Learning and Development*, 18(3):352–376.

Sabrina Horvath, Leslie Rescorla, and Sudha Arunachalam. 2019. The syntactic and semantic features of two-year-olds' verb vocabularies: A comparison of typically developing children and late talkers. *Journal of Child Language*, 46(3):409–432.

Malka Rappaport Hovav and Beth Levin. 2010. Reflections on manner/result complementarity. *Syntax, lexical semantics, and event structure*, pages 21–38.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. Masc: The manually annotated sub-corpus of american english. In *6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40.

Manfred Krifka. 1992. Thematic relations as links between nominal reference and temporal constitution. *Lexical matters*, (24):29.

Beth Levin. 2008. A constraint on verb meanings: Manner/result complementarity. *Cognitive Science Department Colloqium Series, Brown University, Providence, RI, March*, 17:2008.

Beth Levin and Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. *cognition*, 41(1-3):123–151.

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*, abs/1907.11692.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About time: Do transformers learn temporal verbal aspect? In *12th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2022)*, pages 88–101. ACL: Association for Computational Linguistic.

Letitia R Naigles and Erika Hoff-Ginsberg. 1998. Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use. *Journal of child language*, 25(1):95–120.

Hollis S Scarborough. 1990. Index of productive syntax. *Applied psycholinguistics*, 11(1):1–22.

Rico Sennrich. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Marije L Verhage, Carlo Schuengel, Robbie Duschinsky, Marinus H van IJzendoorn, RM Pasco Fearon, Sheri Madigan, Glenn I Roisman, Marian J Bakermans-Kranenburg, and Mirjam Oosterman. 2020. The collaboration on attachment transmission synthesis (cats): A move to the level of individual-participant-data meta-analysis. *Current Directions in Psychological Science*, 29(2):199–206.

Susan Ellis Weismer, Courtney E Venker, Julia L Evans, and Maura Jones Moyle. 2013. Fast mapping in late-talking toddlers. *Applied Psycholinguistics*, 34(1):69–89.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

Appendix

A Instructions to Expert Annotator and Annotation Tool

The instructions that were provided to the expert human annotator before starting the annotation process is shown in Figure 6 and the sample annotation screen is provided in Figure 7. The users were provided with clear definition taken from (Hovav and Levin, 2010; Levin, 2008) paper.

A sample annotation screen is shown in Figure 8. The user can tag the sentences in multiple sessions and there were a total of 200 sentences to annotate. The VerbNet categories are shown on the left.

B LLM Prompting

Figure 9 represents the rule for instructing LLM to focus on the sentence construction while tagging result and manner verbs.

Identifying Manner and Result Verbs in Non-Stative Verbs
Definition: Verbs can be classified into two categories: Non-Stative Verbs and Stative Verbs.

1. Non-Stative Verbs
1.1 Manner Verbs: These verbs lexicalize the manner in which an action/event takes place. *Examples:* cry, hit, pound, run, shout, shovel, smear, sweep, etc.

1.2 Result Verbs: These verbs lexicalize the result or outcome of an event. *Examples:* arrive, clean, come, cover, die, empty, fill, put, remove, etc.

1.2.1 Scalar Result: Describes a change of state in the event, leading to a new final state. *Example:* "John **carved** the wood into a toy."

1.2.2 Scalar Change: Indicates some change of state in the event, even if it does not result in a new final state. *Example:* "John **drove** the car around the parking lot."

2. Stative Verbs
Stative verbs describe a state rather than an action. They are not typically used in the present continuous form.

Examples:
"I don't know the answer." (*I'm not knowing the answer.*) (Ungrammatical)
"She really likes you." (*She's really liking you.*) (Ungrammatical)

Annotation Task:
Your next task is to determine all the applicable categories (from the four listed) that the highlighted verb (in yellow) belongs to in the given sentence. If unsure, mark it as "Not Sure."

Reference Material:
For further understanding, refer to the below PDF (only 2 pages) for insights on manner-result verbs by the original authors.

Figure 6: Guidelines for Identifying Manner and Result Verbs in Non-Stative Verbs

C Qualitative Analysis

Here we illustrate some qualitative cases where, given a sentence as input, we checked the categorization returned by the two models. Both models could identify the distinct nuances between manner and result verbs in most cases. For example, in the sentence "*She sponged the bottle well*" both models correctly classified the verb "*sponged*" as a manner verb, while in the sentence "*She cleaned the bottle well*", both models accurately classified the verb "cleaned" as a result verb. This demonstrates that, irrespective of context, the models developed an understanding of the verb root to distinguish between result and manner connotations.

Additionally, to highlight the capability of the models in distinguishing stative and non-stative verbs, we checked a few sentences. In the sentence "*The mother ran to the market and bought her child a gift, because she loves her a lot*", both models accurately identified the categories of the verbs "*ran*", "*bought*", and "*loves*" as manner, result, and stative, respectively. However, when given the sentence, "*The president learned of a coup plot that might endanger his life*", model 2 incorrectly classified the verb "*endanger*" as stative, while model 1 accurately identified the verb as result.

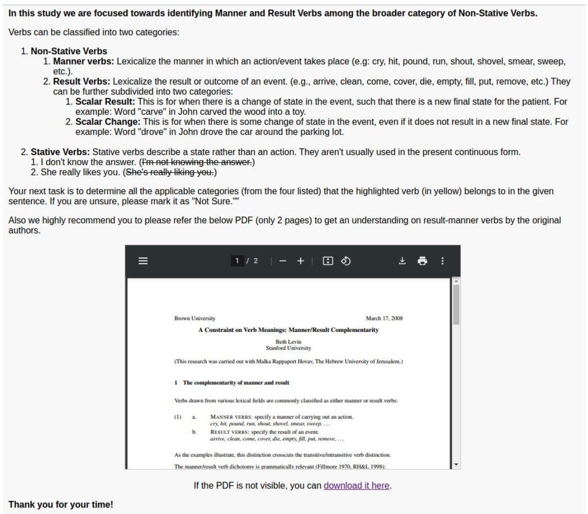

Figure 7: Annotation Screen for Expert Human Annotator.

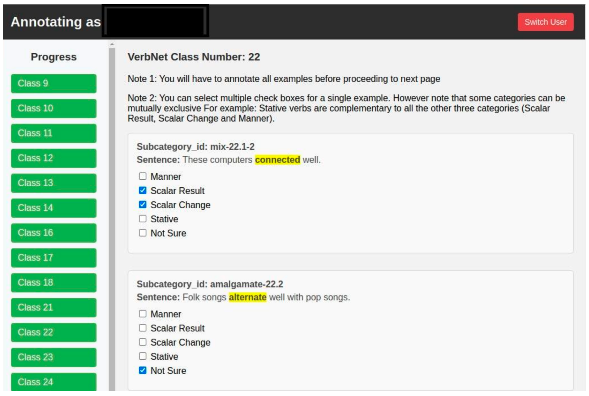

Figure 8: Sample Annotation Screen.

Manner Verbs
Definition: These verbs encode the *how* of an action, focusing on the method or process by which an action is performed rather than its outcome.

Syntactic Diagnostic 1: Unspecified Objects
Manner verbs frequently occur with unspecified or non-subcategorized objects in nonmodal, nonhabitual sentences.
Example: "Anna wept all day." (Acceptable)

Syntactic Diagnostic 2: Causative/Inchoative Alternation
Manner verbs do not participate in the causative/inchoative alternation.
Example: Causative: "John wiped the table."
Inchoative: *"The table wiped."* (Ungrammatical)

Usage:
"She **scribbled** on the notebook." (Focus on the method of writing)

Result Verbs
Definition: These verbs encode the *outcome* or resultant state that follows from an action.

Syntactic Diagnostic 1: Specified Objects
Result verbs typically do not occur with unspecified or non-subcategorized objects. They require a direct object that undergoes a change.

Syntactic Diagnostic 2: Causative/Inchoative Alternation
Result verbs readily participate in the causative/inchoative alternation, appearing both in causative constructions (with an explicit external agent) and in inchoative constructions (where the change occurs spontaneously).
Examples:
Causative: "The child broke the vase." (Agent causes the change)
Inchoative: "The vase broke." (The change occurs without an explicit agent)

Usage:
"He **melted** the ice." (Focus on the resulting state)

Figure 9: Manner vs. Result Verb Sentence Construction Prompt