Unraveling Verb Root Semantics: A Scalable Approach to Manner and Result Categorization

Anonymous ACL submission

Abstract

In this paper, we address the computational identification and categorization of verbs into result, and manner verbs—a distinction that has been shown to influence child vocabulary acquisition and later difficulties with language learning such as Developmental Language Disorder (DLD). Within this framework, manner verbs encode the dynamic “how” of an action, and *result* verbs, denote a change in outcome. Prior work has been limited to a narrow subset of VerbNet, and relied mainly on human linguistic reasoning without scalable computational methods. In contrast, we leverage Large Language Models (LLMs) as an expert annotator to generate synthetic annotations on 436 out of 487 VerbNet classes over sentences drawn from MASC and InterCorp dataset. These annotations serve as training data for a RoBERTa-based classifier, which achieves an accuracy of 89.6% overall on gold annotated datasets. To the best of our knowledge, this work presents the first large-scale computational approach to result and manner verb classification.

1 Introduction

Research on the development and representation of natural language concepts has traditionally focused on noun-based phenomena, in part because children’s early vocabularies are dominated by concrete nouns (Gentner, 1982; Gentner and Boroditsky, 2001). However, recent studies suggest that the number and types of verbs (Behrend, 1990) produced by children—especially during the critical period around age two—are better predictors of later grammatical skills and language disorders (Toddlers’ Verb Lexicon Diversity and Grammatical Outcomes). Unlike nouns, which refer to tangible objects, verbs encode actions and events that often involve complex semantic and syntactic structures. Just as some nouns are learned earlier i.e.. those denoting basic level categories e.g. dog than later i.e. those describing superordinate categories e.g. mammal, different categories of verbs also vary in their prevalence in early vocabularies. One semantic distinction that has been found to be relevant in previous research is the distinction between *manner verbs* (e.g., *nibble, rub, scribble, sweep, flutter, laugh, run, swim*) which incorporate information about the execution or "how" of an action, and *result verbs* (e.g., *clean, guillotine, bake, climb, cover, empty, fill, freeze, melt, open, arrive, die, enter, faint*) which emphasize the "what" of an action (Hovav and Levin, 2010; Levin, 2008). Recent findings by Horvath et al. (2022) indicate that the relative proportions of manner and result verbs in children’s speech are statistically significant predictors of late language disorders, differentiating between late talkers and typically developing children. Moreover, studies suggest that children who produce more manner verbs also produce more verbs overall. (Horvath et al., 2019, 2022) Examining the relative frequency of verb types as a predictor of later language development is particularly important, as many as half of children with early signs of language delays go on to have Developmental Language Disorder, but the predictors of which students will ultimately have language outcomes in the typical range and those will have further clinical diagnoses are not well understood. While there has been considerable work on grammatical aspects such as parts-of-speech tagging (DeRose, 1988), semantic categorizations based on verb root meaning remain underexplored in computational linguistics. Recent efforts to classify similar lexical properties like telicity (although a compositionally determined category) (Friedrich et al., 2022; Friedrich and Gateva, 2017) have shown both the benefits of leveraging computational approaches for such categorization while also reflecting the inherent challenges of capturing fine-grained semantic distinctions. For instance, despite parts-of-speech models achieving accuracies over 97% (Manning, 2011), even the best models for telicity classification reach only 87% (Metheniti et al., 2022; Friedrich et al., 2022), with human annotators achieving about 79% (Friedrich et al., 2016). A major barrier to progress in result and manner verb classification (unlike Situation Entity types like telicity, durativity, and stativity etc.) is the absence of gold-standard annotated datasets. To bridge this gap, we propose a scalable approach that leverages LLMs as expert annotators. By utilizing established result and manner definitions and a small set of illustrative examples, we prompt LLMs to annotate the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008), and InterCorp parallel corpus (ek Čermák and Rosen, 2012) datasets at a sentence level. Current gold-standard annotations for result and manner tags span 151 out of 487 VerbNet (Brown et al., 2019) classes, while we generate annotations for 436 VerbNet classes (Kipper et al., 2008). We then fine-tune a pretrained RoBERTa-based (Liu et al., 2019) classifier on these LLM-defined annotations and demonstrate that it can effectively distinguish between manner and result verbs, achieving an accuracy of 91.8% on 160 existing gold-standard annotations, and 86% on another 110 annotations provided by an expert annotator.

In summary, our contributions are:

- We introduce the first scalable, computational framework for identifying and classifying *result* and *manner* verbs, given any sentence “from-the-wild”.
- We propose a methodology that leverages LLMs to generate training data in the absence of large-scale gold-standard annotations.
- We will publicly release our code and annotated dataset covering almost 90% of the entire VerbNet hierarchy, for future research.

To navigate this paper, Section 1 discusses the motivation for the research in the context of past related works from the developmental, linguistic and computational sciences perspectives. And with this being a new problem in the computational linguistics domain, we dedicate Section 2 to defining the constructs-of-interest (manner and result verbs) using several examples for illustration. Section 3 presents the logical steps an expert would use in distinguishing between the two constructs and describes how we used these to create LLM prompts to annotate our training data. In Section 4, we describe the end-to-end process including training data annotation, model training, and test data creation. Section 5 discusses the implementation details of the computational model, Section 6 presents our experimental results (from inference), and we conclude the paper and discuss our future work in Section 7.

2 Understanding Verb Root Meaning

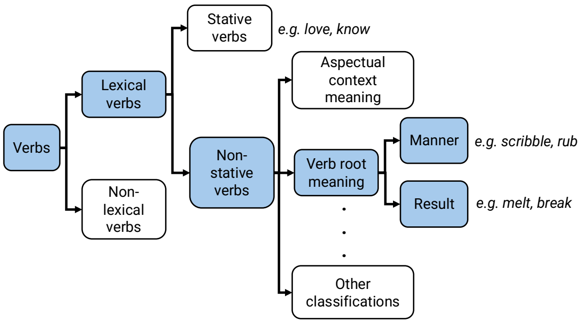


Figure 1: Hierarchy of verb classification, with manner and result verbs as subdivisions of non-stative verbs

Figure 1 shows 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* the change of outcomes that occurs, without specifying the manner in which the action was performed (e.g., *clean, melt, fill, arrive*).

Unlike classifications such as telicity, which are determined at clause level (Friedrich and Gateva, 2017) the manner/result distinction is **inherent to the verb root** (Levin, 2008), implying that a verb's classification remains stable regardless of argument selection or context.

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—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 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 provide useful cues, as manner and result verbs tend to appear 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 drier dried the clothes for two hours* (Atelic: no clear endpoint)
- *The drier dried the clothes in two hours* (Telic: the drying is completed)

Since telicity is observed to be influenced by syntax (temporal adverbial choice), we do not use it as rule for classifying manner/result verbs.

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 process including (i) annotation of training data for manner/result tags (ii) training of tagging model and (iii) the creative ways in which we obtained near gold-standard test data.

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);

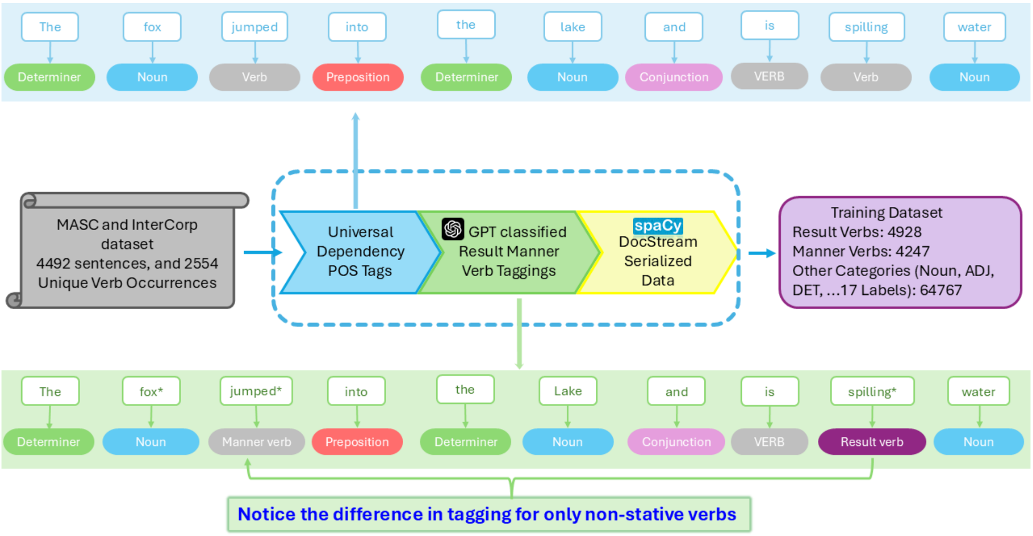

The
fox
jumped
into
the
lake
and
is
spilling
water
Determiner
Noun
Verb
Preposition
Determiner
Noun
Conjunction
VERB
Verb
Noun
MASC and InterCorp dataset 4492 sentences, and 2554 Unique Verb Occurrences
Universal Dependency POS Tags
GPT classified Result Manner Verb Taggings
spaCy DocStream Serialized Data
Training Dataset Result Verbs: 4928 Manner Verbs: 4247 Other Categories (Noun, ADJ, DET, ...17 Labels): 64767
The
fox*
jumped*
into
the
Lake
and
is
spilling*
water
Determiner
Noun
Manner verb
Preposition
Determiner
Noun
Conjunction
VERB
Result verb
Noun
Notice the difference in tagging for only non-stative verbs

Figure 2: Overview of our data generation pipeline.

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.

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

Our initial gold-standard dataset consisting of 83 verbs (34 result verbs and 49 manner verbs), was obtained by manually combing through the linguistics literature Levin (2008); Hovav and Levin (2010); Beavers and Koontz-Garboden (2012); Levin and Hovav (1991) for works describing the lexical semantics, and verb-root classification. We refer to this as the *Linguists verb-root* data.

The next annotations we leverage is from Horvath et al. (2022) Psycholinguists paper where they tagged 36 Result Verbs and 41 Manner Verbs from the MacArthur-Bates Communicative Development Inventory (MBCDI)[1]. We refer to this as the manner/result tagged set from MBCDI as the *Psycholinguistic verb-root* data. Since the above two datasets covered only 151 out of 487 VerbNet classes, we collaborated with an expert linguist specializing in this area to obtain additional annotations across a broader range of verb types. This was necessary because VerbNet groups verbs based on semantic structure, and we aimed to evaluate our model on a diverse set of verb roots. Guided by VerbNet, we constructed 200 new sentences, expanding the coverage to 346 VerbNet classes. The expert linguist then annotated the verbs, categorizing 23 as stative, 48 as result, 62 as manner, and 67 as 'unsure.' Details of the instructions provided to the expert annotator and the design of annotation tool are discussed in the Appendix A. We refer to this as the *Expert-annotated verb-root* data.

We evaluate our models on unseen data using annotations from the three gold-standard datasets. Notably, these datasets are separate from the MASC corpus, which we labeled to get the training set, thereby ensuring that these 3 new datasets serve as out-of-training samples.

5 Computational Modeling

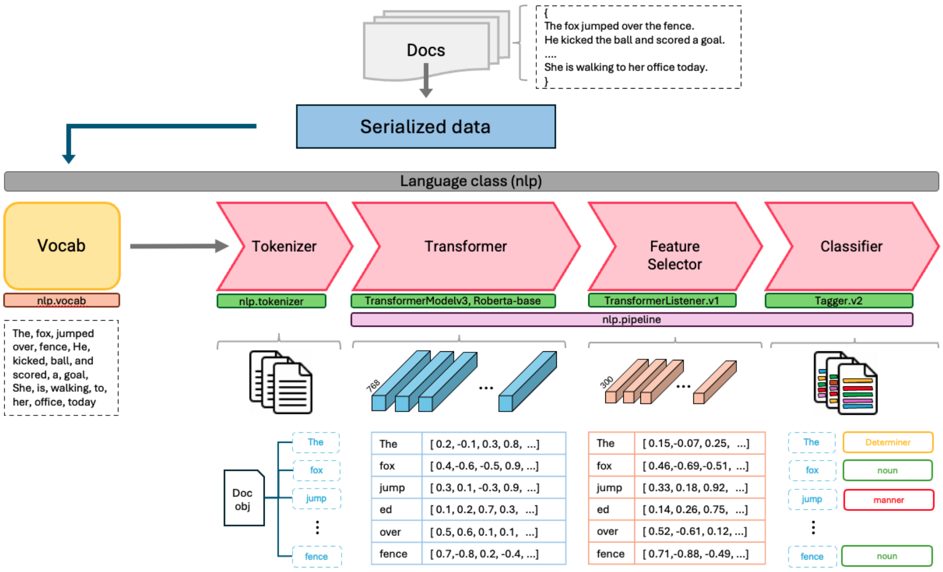
This section outlines our computational approach for classifying verbs according to both *manner/result* and *stative/non-stative* properties.

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

[1] MBCDI is a well-established and extensively studied first-language assessment tool designed to evaluate children's lexical development in English as their first language.


Docs
{
The fox jumped over the fence.
He kicked the ball and scored a goal.
....
She is walking to her office today.
}
Serialized data
Language class (nlp)
Vocab
nlp.vocab
The, fox, jumped over, fence, He, kicked, ball, and scored, a, goal, She, is, walking, to, her, office, today
Tokenizer
nlp.tokenizer
Transformer
TransformerModelv3, Roberta-base
Feature Selector
TransformerListener.v1
Classifier
Tagger.v2
nlp.pipeline
768
300
Doc obj
The
fox
jump
fence

The	[0.2, -0.1, 0.3, 0.8, ...]
fox	[0.4,-0.6, -0.5, 0.9, ...]
jump	[0.3, 0.1, -0.3, 0.9, ...]
ed	[0.1, 0.2, 0.7, 0.3, ...]
over	[0.5, 0.6, 0.1, 0.1, ...]
fence	[0.7,-0.8, 0.2, -0.4, ...]

The	[0.15,-0.07, 0.25, ...]
fox	[0.46,-0.69,-0.51, ...]
jump	[0.33, 0.18, 0.92, ...]
ed	[0.14, 0.26, 0.75, ...]
over	[0.52, -0.61, 0.12, ...]
fence	[0.71,-0.88, -0.49, ...]

The	Determiner
fox	noun
jump	manner
⋮	
fence	noun

Figure 5: Overview of model architecture

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 syntactic 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 $= 128$.

Schedule and early stopping. 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.

Implementation. 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.

6 Experiments & 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

	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.

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. We empirically observed that this dataset contained a number of activity such as *paint*,*dump*, *drink*, etc. These verbs appear to have a manner connotation, but the dataset classified them as result. This suggests that our model is performing to our expectation based on the expert-guided governing rules we provided. 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 Conclusion

We have presented a novel computational framework for categorizing verbs based on their event structure. By leveraging annotations generated via Large Language Models (LLMs), our approach is trained to distinguish between manner and result verbs, expanding the number of unique annotated VerbNet classes from 151 to 436. Our methodology integrates rigorous linguistic diagnostics including syntactic tests (object omission and causative/inchoative alternation) and semantic cues (scalar versus non-scalar changes) into the training data annotation pipeline. We then apply a RoBERTa-based model for simultaneous POS and verb-root classification.

Experimental evaluation demonstrates strong performance, achieving up to 89.6% average accuracy across three gold-standard datasets. Notably, our results indicate that the inherent semantic properties of non-stative verb roots are more important for accurate classification than properties related to the sentence structure only. These findings highlight the fact that if designed properly, computational models such as we propose, can effectively understand the distinctions between manner and result verbs.

In our future work, we plan to incorporate diverse linguistic data to mitigate any bias in the verb classification task. We also plan to extend this framework to additional languages and explore further integration of syntactic and semantic features, reaffirming the potential of combining linguistic theory with deep learning methods for advanced language understanding. In collaboration with our Language Development co-authors, we plan to apply the model to an existing large text corpus to predict for Developmental Language Disorder (DLD) in children. This could poses some potential risk in that we will be using the outputs of the model to partly determine the extent of interventions provided to children.

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-stativ 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 Impacts

The following section discusses some of the potential ethical impacts of the proposed work:

- LLMs are trained on large text corpora and may inherit linguistic and cultural biases which could potentially result in overlooking any linguistic diversity; people from different geographic locations or demographics often speak English with their own distinct vocabulary, pronunciation, and grammar patterns. So an LLM possessing Anglocentric biases may result in the marginalization of non-traditional, non-standard Anglo-English speakers.

- Because this verb categorization model will be used in clinical, and diagnostic settings, i.e. for detecting language impairments or assessing linguistic proficiency in children, there is a risk of misclassification due to the model's reliance on statistical patterns rather than cognitive or developmental principles.

- The proposed model runs an accessibility and exclusion risk as far as "who benefits from the research". Since the model is primarily trained on Anglocentric data only, it may exclude communities whose linguistic patterns are not well-documented, such as African America children who might speak with a distinct dialect (spoken by many African Americans) with its own unique grammatical features.

Broader Impacts

The broader impacts of this research extend across multiple domains, including linguistics, AI, developmental psychology, and society. Below are some key areas:

- Late Talkers make up 9-20% of children between 18-30 months of age, and exhibit slower expressive language development relative to their peers. Approximately half such late talkers will be diagnosed with persistent language deficits, such as Developmental Language Disorder (DLD), which can have far-reaching consequences (e.g. lower educational outcomes, (Catts et al., 2012); increased risk for unemployment, (Conti-Ramsden et al., 2018)). Hence, identifying predictors of later language disorders can enable clinicians to provide timely intervention for those who need it and have the potential to address the persistent inequities in access to speech-language services

- The proposed methodology plays a crucial role in AI modeling by enabling the integration of domain expertise. It facilitates the design of LLM prompts from governing rules from learned from domain experts, where such prompts can be used to generate annotated data useful for subsequent model training. This approach significantly enhances future AI models by embedding expert knowledge into their development, improving both their accuracy and domain relevance.

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.

Douglas A Behrend. 1990. The development of verb concepts: Children's use of verbs to label familiar and novel events. *Child Development*, 61(3):681–696.

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.

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. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language*, 2:301–334.

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

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*.

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/-manfred krifka. *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.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

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.

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

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.

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.

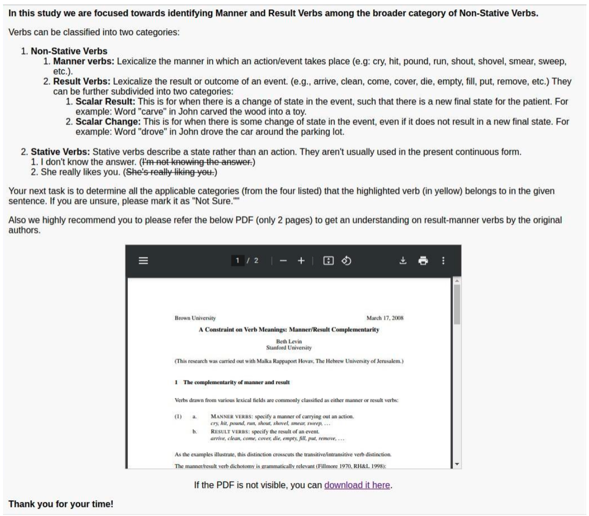
Examples:
"I don't know the answer." (*I'm not knowing the answer.*) (Ungrammatical)
"She really likes you." (*She's really liking you.*) (Ungrammatical)

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

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.



In this study we are focused towards identifying Manner and Result Verbs among the broader category of Non-Stative Verbs.

Verbs can be classified into two categories:

1. **Non-Stative Verbs**
    1. **Manner verbs:** Lexicalize the manner in which an action/event takes place (e.g: cry, hit, pound, run, shout, shovel, smear, sweep, etc.).
    2. **Result Verbs:** Lexicalize the result or outcome of an event. (e.g., arrive, clean, come, cover, die, empty, fill, put, remove, etc.) They can be further subdivided into two categories:
        1. **Scalar Result:** This is for when there is a change of state in the event, such that there is a new final state for the patient. For example: Word "carve" in John carved the wood into a toy.
        2. **Scalar Change:** This is for when there is some change of state in the event, even if it does not result in a new final state. For example: Word "drove" in John drove the car around the parking lot.

2. **Stative Verbs:** Stative verbs describe a state rather than an action. They aren't usually used in the present continuous form.
    1. I don't know the answer. (~~I'm not knowing the answer.~~)
    2. She really likes you. (~~She's really liking you.~~)

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 you are unsure, please mark it as "Not Sure.""

Also we highly recommend you to please refer the below PDF (only 2 pages) to get an understanding on result-manner verbs by the original authors.

Brown University March 17, 2008

A Constraint on Verb Meanings: Manner/Result Complementarity

Beth Levin
Stanford University

(This research was carried out with Malka Rappaport Hovav, The Hebrew University of Jerusalem.)

1 The complementarity of manner and result

Verbs drawn from various lexical fields are commonly classified as either manner or result verbs:

(1) a. MANNER VERBS: specify a manner of carrying out an action.
cry, hit, pound, run, shout, shovel, smear, sweep, …
b. RESULT VERBS: specify the result of an event.
arrive, clean, come, cover, die, empty, fill, put, remove, …

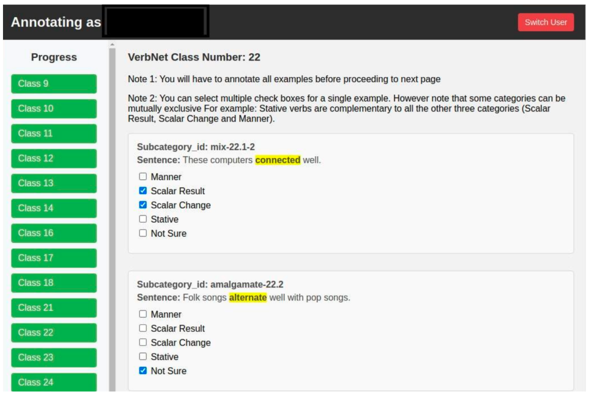
As the examples illustrate, this distinction crosscuts the transitive/intransitive verb distinction.

The manner/result verb dichotomy is grammatically relevant (Fillmore 1970, RH&L 1998):

If the PDF is not visible, you can download it here.

Thank you for your time!

Figure 7: Annotation Screen for Expert Human Annotator.



Annotating as [redacted]

VerbNet Class Number: 22

Note 1: You will have to annotate all examples before proceeding to next page

Note 2: You can select multiple check boxes for a single example. However note that some categories can be mutually exclusive For example: Stative verbs are complementary to all the other three categories (Scalar Result, Scalar Change and Manner).

Subcategory_id: mix-22.1-2
Sentence: These computers **connected** well.
☐ Manner
☑ Scalar Result
☑ Scalar Change
☐ Stative
☐ Not Sure

Subcategory_id: amalgamate-22.2
Sentence: Folk songs **alternate** well with pop songs.
☐ Manner
☐ Scalar Result
☐ Scalar Change
☐ Stative
☑ Not Sure

Figure 8: Sample Annotation Screen.

B LLM Prompting

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

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

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.