Utrecht University

MASTER'S THESIS

Resultatives in word embeddings

Author:
Jan Anthonie de Groot, BSc.

Student Number:
6268188

Supervisor:
dr. Tejaswini Deoskar

Second reader:
dr. Rick Nouwen

A thesis submitted in fulfilment of the requirements
for the degree of Master of Science

in

Artificial Intelligence

at the

Faculty of Science

13 November 2024

Abstract

Jan Anthonie de Groot, BSc.

Resultatives in word embeddings

Verbs can be categorised into classes based on the syntactic frames and alternations they can and cannot participate in (Levin, 1993). This study investigates whether pre-trained language model (PLM) word embeddings can be used to determine if a verb participates in the resultative construction. This construction describes that a change of state has taken place as the result of an action (Goldberg and Jackendoff, 2004; Levin, 1993). In this thesis, we extend the lexical dataset LaVA presented in Kann et al. (2019), to include the resultative construction, adding 70 new verbs, and extending the annotations of the existing verbs in the dataset. We further train a logistic regression classifier on the verb embeddings from BERT (Devlin et al., 2019) to determine which verbs participate in which frames. Our analysis shows that the performance on resultative frames using static pre-trained embeddings is consistent with similar works.

Keywords: Lexical semantics, resultative construction, pre-trained embeddings, verb alternations, natural language processing, artificial intelligence

Contents

1 Introduction

1.1 Verb alternations

Humans have implicit knowledge of the syntax of the (native) language(s) they speak, and are able to recognise small nuances between various sentences. These are integral to a person's linguistic capabilities. Verbs are a significant component of syntax, and serve as a good illustration of this innate knowledge (Levin, 1993, p. 2).

Verbs take arguments and those arguments can appear in various positions. Speakers of English intuitively know in which positions verbs and their arguments may be placed. For example, the verb *give* may realise its arguments in two different ways (Kann et al., 2019; Levin, 1993):

(1) a. Liz gave a gift to the boy.
b. Liz gave the boy a gift.

The alternation in the structure of a verb's arguments is called a *verb alternation*, while each of the variants is called a *frame*. The alternation above is an example of the Dative alternation (Levin, 1993, pp. 45–48; Jurafsky and Martin, 2024a, pp. 3–4).

Verbs can be categorised into classes based on the syntactic alternations they can and cannot participate in. In the following example, showing what is known as the Causative-Inchoative alternation (Levin, 1993; Piñón, 2001; Schäfer, 2009), the verb *break* can be either transitive (2a) or intransitive (2b). It is said to *select* multiple related frames: the sentences have different argument structures, but both are semantically valid. However, two verbs that intuitively have similar semantics may differ in the syntactic frames they can appear in. Unlike *break*, *hit* does not produce two valid sentences, as shown in example (3). Whether or not a verb can participate in a certain frame is referred to as the *frame-selectional property* of the verb (Jurafsky and Martin, 2024a; Levin, 1993; Levin and Rappaport Hovav, 1995; Yi et al., 2022b).

(2) a. Jessica broke the window.
subject (transitive) verb direct object
b. The window broke.
subject (intransitive) verb

(3) a. Jessica hit the window.
b. * The window hit.

Although much has been written on verb alternations from a linguistics point of view (e.g. Arad, 2006; Dikken and Hoekstra, 1994; Fillmore, 1970; Levin, 1993; Piñón, 2001; Simpson, 1983), computational approaches have appeared in the literature relatively recently.

Alternation	Verb Frames	Example Sentences	
Causative-Inchoative	Causative	Jessica dropped the vase.	Jessica blew the bubble.
	Inchoative	The vase dropped.	* The bubble blew.
Dative	Preposition	Liz gave a gift to the boy.	Liz administered a test to the kid.
	Double-Object	Liz gave the boy a gift.	* Liz administered the kid a test.
Spray-Load	*with*	Sue loaded the truck with wood.	Sue coated the deck with paint.
	Locative	Sue loaded wood onto the truck.	* Sue coated paint on the deck.
there-Insertion	No-There	Fear remained in my mind.	A girl focused on the quiz.
	There	There remained fear in my mind.	* There focused on the quiz a girl.
Understood-Object	Reflexive	Ada clapped her hands.	Ada permed her hair.
	Non-Reflexive	Ada clapped.	* Ada permed.

Table 1.1: Example sentences for each frame of the 5 alternations in Kann et al. (2019). For each alternation, the third column shows instances of alternating verbs and the fourth column shows non-alternating verbs.

There have been several studies on the automatic identification of verb alternations. Often, these studies focus on a particular alternation, considering that there is a vast assortment of them. To the best of our knowledge, predicting verb class membership based on embeddings was first attempted using an artificial neural network (Kann et al., 2019), utilising the CoLA framework developed by Warstadt et al. (2019). Concurrently, the use of recurrent neural networks was investigated by Seyffarth (2019), and later by Loáiciga et al. (2021). Other techniques, such as support vector machines (Seyffarth and Kallmeyer, 2020) and linear regression (Yi et al., 2022b) were also trialled.

This thesis is influenced by the work of Kann et al. (2019), as well as by Yi et al. (2022b). Both studies make use of datasets developed to test if artificial neural networks (ANNs) can correctly distinguish acceptable from unacceptable verb–frame combinations. The data is partially sourced from Levin (1993), which lists the semantic classes of over 3100 English verbs and the alternations they participate in. This book will be discussed in more detail in section 2.3. Like examples (2) and (3) for Causative-Inchoative above, each alternation consists of two different syntactic frames, which vary in the number and/or order of arguments they can take. Five of the syntactic verb frame alternations provided by Levin (1993) are used in LaVA, a lexical dataset developed by Kann et al. (2019), shown in Table 1.1. For each of the alternations, LaVA gives participation judgements for 516 verbs. Section 2.1 describes the datasets in more detail.

So far, only these five alternations have been tested. The results point to the cautious confirmation that language models can indeed learn to correctly classify verb–frame combinations. However, these five are among the most common alternations with the greatest range of verbs. Four of the five alternations used are in the top 10 when sorted by number of verbs listed in Levin (1993).[1]

This raises the question if language models are able to predict verb–frame combinations for rare(r) alternations to the same degree as they can for common alternations. With that question in mind, we turn to the resultative construction. Although a great variety of verbs can participate in the resultative construction, few verbs are actually listed in Levin (1993). Building a dataset with participation judgements for verbs in the construction could support further research into the representation of verb frames in embeddings.

[1] Based on Lawler (n.d.), we know the number of verbs listed at each alternation in Levin (1993) Part I. Sorting all alternations by number of verbs gives the following ranking for the alternations used in LaVA. 1. Causative-Inchoative: 529, 2. Dative: 336, 6. *there*-Insertion: 224, 10. Spray-Load: 214, 25. Understood-Object: 110, 52. Resultative Construction: 22.

The *resultative construction* describes that a change of state has taken place as the result of an action (Levin, 1993, p. 101; Goldberg and Jackendoff, 2004). This is best illustrated with an example.

(4) The waiter wiped the table **clean**.

(5) The rooster crowed me **awake**.

In sentence (4), the adjective *clean* (the result) describes the resulting state of *the table* after *the waiter wiped*. This is an example of a *selected NP resultative*, because the postverbal NP (here: *the table*) is the object of the verb if the result were to be omitted. Sentence (5) does not show this behaviour, and as such, is an example of a *non-selected NP resultative*. Section 2.2 provides an in-depth description of the resultative construction and its variants.

1.2 Research questions

Based on the shortcomings in the previous work above and our plan to address the limited number of alternations in LaVA, we arrive at the following research question.

To what extent are selected NP resultatives and non-selected NP resultatives represented in word embeddings?

In other words, this thesis aims to find out if a verb's word embedding can be used to predict if that verb can appear in the resultative construction. A number of sub-tasks need to be completed before we can solve the main research question. Each of these is listed below in italics, directly followed by a description.

1. *Extend LaVA to accommodate data for the new frames.*
   In its current state, LaVA contains data for five alternations. As we want to make a similar dataset for a new alternation, we need to allocate space for the new frames, and assign each verb a label indicating the verb's participation in the frames.

2. *Extend LaVA to include more verbs that may participate in the construction.*
   The verbs in the dataset were chosen such that each would participate in at least one frame of one alternation. Each frame has more than 100 verbs with either a positive or negative label. In order for the resultative to be as representative as the other five alternations, the number of verbs must increase.

3. *Run an experiment using the extended dataset.*
   Having built the dataset, we can now test if word embeddings can distinguish selected NP resultatives and non-selected NP resultatives.

Kann et al. (2019) demonstrate that verb embeddings can distinguish between syntactic frames. Yi et al. (2022b) find that the embeddings do encode information about the five alternations in LaVA. Although the resultative construction behaves differently, we hypothesise that verb embeddings are still useful in classifying verb–frame combinations. Our results show roughly equal metrics for the existing and new frames, suggesting that there is a comparable level of representation in verb embeddings. However, overall performance on the full dataset is worse than on the original LaVA.

1.3 Relevance

The frames of a verb alternation often exhibit the same semantics, whereas the syntax is altered. The syntactic structures may have different interpretations for non-alternating verbs. Knowing which verbs participate in which alternations could potentially improve semantic role labelling tasks, where each predicate in a sentence automatically gets assigned a semantic role. Frame induction could also benefit from this knowledge, for the goal of this task is to group predicates (verbs, usually) together based on the semantic frames evoked by that verb. Advances in these fields may in turn yield improvements in tasks such as information extraction and question answering, among others (Gildea and Jurafsky, 2002; Jurafsky and Martin, 2024a; Seyffarth and Kallmeyer, 2020).

The results from the experiments by Yi et al. (2022b) seem to point to the notion that embeddings encode information about verb alternation classes, this has not been replicated in any research since. By running a very similar experiment, we may obtain new insights that support or contradict their claims. Replicability is an essential part of science, and repeating the experiment acts as some sort of quality control (National Academies of Sciences, Engineering, and Medicine et al., 2019).

1.4 Outline

The thesis is organised as follows: Chapter 2 introduces the necessary context and concepts, and gives a description of the resultative construction. In Chapter 3, we give a detailed account of our approach to creating a dataset. Once the dataset is complete, we are interested in how it fares. We give an overview of the experiments that ours draws inspiration from in Chapter 4 and describe the implementation of the experiment in Chapter 5. This chapter also reports the results. In Chapter 6 we discuss the implication of the results and recommend future research. Chapter 7 concludes the study.

2 Theoretical background

In this part, we will explore the concepts that recur throughout the thesis. We introduce the datasets we are working with in section 2.1, followed by section 2.2, in which we discuss the resultative construction, a phenomenon not present in LaVA or FAVA. We conclude with a description of Levin (1993).

2.1 Existing work

Since the thesis is inspired by and built upon the works of Kann et al. (2019) and Yi et al. (2022b), some knowledge about these papers is required to understand this project.

In order to assess whether LMs can distinguish acceptable from unacceptable verb–frame pairs based on word or sentence embeddings, Kann et al. (2019) created two datasets: LaVA, a lexical corpus, and FAVA, an acceptability judgement sentence corpus. Section 2.1.1 provides a description on LaVA. Section 2.1.2 is a rather extensive description of FAVA, even though we did not do any substantial work in that direction. However, we feel that being aware of the dataset and how it was created, helps understand some of the choices we made.

2.1.1 Description of LaVA — Lexical Corpus

The **L**exic**a**l **V**erb–frame **A**lternations (LaVA) dataset is constructed from 516 verbs taken from the verbs listed at each of the five alternations (see Table 1.1) in Levin (1993). The dataset lists whether each verb participates in 10 classes[2] corresponding to verb frames, denoted as follows: '1' for membership, '0' for non-membership, and 'x' where membership is unknown. Table 2.1 shows some entries for verbs from the example sentences in the Introduction (1.1).

verb	sl	sl_noloc	sl_nowith	inch	non_inch	there	non_there	dat_both	dative_to	dat_do	refl_op	refl_only
broke	0	0	0	1	0	x	x	0	0	0	0	1
dropped	0	0	1	1	0	x	x	0	0	0	0	0
gave	0	0	0	0	x	0	x	1	0	0	0	0
hit	x	x	x	0	1	x	x	0	0	0	0	0
loaded	1	0	0	x	x	x	x	0	0	0	0	0

TABLE 2.1: Sample of the LaVA dataset as published by Kann et al. (2018). A glossary of the column names is located in Appendix A.

Sometimes, no negative examples can be obtained, because they do not exist. For instance, while some verbs may appear in the No-There frame and not in the There frame, there are no English verbs that can appear only in the There frame. For a similar reason,

[2] In alternations like DATIVE, verbs can take 3 arguments, whereas verbs alternating in the *there*-INSERTION only take 2 (Alexiadou and Schäfer, 2011). Where applicable, the alternations are split into three verb frames, resulting in 12 classes. These can be combined such that one can obtain meaningful data about *two* frames (Kann et al., 2019; Yi et al., 2022a).

LaVA contains no verbs that appear in the Inchoative, but not in the Causative frame. As a result, word-level classifications for these frames are trivial (Kann et al., 2019, p. 290).

2.1.2 FAVA — Acceptability Judgements Corpus

The **F**rames and **A**lternations of **V**erbs **A**cceptability dataset (FAVA) is created to investigate grammaticality judgement at the sentence level. The sentences are generated such that the main verb is the only factor in its grammaticality, exploiting the fact that a verb's frame-selection affects grammaticality of the entire sentence. If the main verb can participate in the alternation according to LaVA, then both sentences will be grammatical, whereas one of the sentences will be ungrammatical if the main verb cannot.

FAVA consists of 9413 semi-automatically generated sentences partially formed from the verbs in LaVA, along with grammaticality judgements obtained from a combination of the annotations in LaVA and their own. Here, too, does '1' denote acceptable sentences and '0' unacceptable sentences. Even though the paper states that FAVA is constructed from the lexical corpus, i.e. LaVA (Kann et al., 2019, pp. 290–291), the actual dataset also uses verbs that do *not* occur in LaVA.

The verbs in the lexical sets are chosen based on their similar frame properties and other components to build a sentence. Sentences with various syntactic frames are automatically generated with these sets. For example, lexical set (6) generates $2 \cdot 3 \cdot 3 \cdot 1 = 18$ minimal pairs of sentences as in (7).

(6) verbs = {hung, draped}
patients = {the blanket, the towel, the cloth}
locations = {the bed, the armchair, the couch}
prepositions = {over}

(7) a. Betty draped the blanket over the couch.
b. * Betty draped the couch with the blanket.

Kann et al. (2019) construct a sentence dataset for each of the five verb alternations in Table 1.1. A few examples of sentences in FAVA are shown below. Each row starts with the alternation the sentence is part of, followed by the grammaticality judgement, and finally the sentence itself.

```
dat     1    jason leased a car to the tenant .
dat     1    jason leased the tenant a car .
dat     0    jason tipped 20 pounds to rebecca .
inch    1    rebecca steered the car .
inch    1    the car steered .
inch    0    the onion sliced .
```

2.2 Resultative Construction

The *resultative phrase* describes the state achieved by the noun phrase as a result of the verb (Goldberg and Jackendoff, 2004; Levin, 1993; Simpson, 1983). In this thesis, we will use the term 'resultative sentence' to indicate a full sentence including a subject, a verb and a resultative phrase. The terms 'result', 'resultative' and 'resultative phrase' may be used interchangeably.

Consider resultative sentence (4) once again, repeated here for reference:

(4)	The waiter	wiped	the table	**clean**.
	subject	verb	direct object	resultative

In this sentence, the waiter wiped the table, and the table became clean as a result of the wiping. The resultative, *clean*, is predicated of the direct object, *the table* (Levin, 1993). Examples (8) and (9) show two more sentences where the direct object is 'modified' by the resultative (Levin, 1993; Richards, 2017).

(8) Jasmine pushed the door **open**.

(9) The dog poked me **awake**.

We can see that modifying the subject (for example) does not hold. Example (10) cannot be interpreted such that Polly becomes dirty by cooking, nor does the arriving make Willa breathless in (11). While it is clear what the intended meanings of these sentences are, neither of these are resultative (Levin, 1993).

(10) * Polly cooked the cookies **dirty**.

(11) * Willa arrived **breathless**.

So far, we have looked at sentences where the postverbal NP (the NP directly following the verb) is selected by the verb. These are called *selected NP resultatives* (Levin, 2015). This selection becomes apparent when the result is omitted. By removing the result from examples (4), (8) and (9), we obtain examples (12)–(14). The sentences are still syntactically and semantically valid without the result:

(12) The waiter wiped the table.

(13) Jasmine pushed the door.

(14) The dog poked me.

However, the verb does not always select the NP in resultative sentences. In these *non-selected NP resultatives* (Levin, 2015), one cannot simply remove the result: doing so will produce an ungrammatical sentence. The following examples show this well (Levin, 2015, 2017).

(15) a. The rooster crowed me **awake**.
b. * The rooster crowed me.

(16) a. The maid poured the cup **full**.
b. * The maid poured the cup.

Sentences with a resultative phrase can take various forms; both transitive and intransitive structures show up in corpora (e.g. Boas, 2003). They are not restricted to 'active' verbs either: resultatives can be predicated of subjects of unaccusative or passive verbs, as demonstrated by (17) and (18), respectively (Levin, 1993, p. 100).

(17) The river froze **solid**.

(18) The metal was hammered **flat**.

Selected NP resultatives are only found with transitive verbs. Sometimes the NP might be a reflexive pronoun (Levin, 2015, p. 2). A resultative sentence can assume various syntactic structures (Levin, 1993); a selection of which is itemised here.

- subject + transitive verb + object + resultative phrase (The waiter wiped the table clean)
- subject + intransitive verb + resultative phrase (The river froze solid)
- subject + transitive verb (past simple passive) + resultative phrase (The metal was hammered flat)
- subject + transitive verb + reflexive NP + resultative phrase (She scrubbed herself red)

Unlike the five constructions used in LaVA so far, the resultative construction is not a verb alternation in the sense that the order of the verb's arguments changes. Yet, this is not an issue, as we can consider the selected NP resultative and the non-selected NP resultative as two distinct frames. Technically, we are deviating somewhat from the definition of a 'frame' as presented in section 1.1. We will ignore this in the rest of the thesis.

2.3 English Verb Classes and Alternations: A Preliminary Investigation — Levin (1993)

This thesis relies heavily on the data presented in Levin (1993), so it is necessary to understand how it is structured and thus how we can use it for our purposes.

The book is divided into two parts, encompassing 191 classes and 80 alternations. Part I describes the verb alternations, such as the ones described in section 1.1. Each section is dedicated to a specific alternation, and usually starts by listing the alternating verbs, i.e. the verbs for which both frames form a grammatical sentence. Often, a list of verbs that only allow one frame is given, along with example sentences.

Part II is an extensive description of verb classes whose members behave similarly across different alternations (e.g. Verbs of Putting: *arrange*, *place*, *put*, *set* (Levin, 1993, p. 112)). For each verb class, its members (the verbs) and properties (the alternations) are listed along with example sentences. For most verb classes, the properties include alternations that it does *not* participate in. The following excerpt from section 10.3 (Levin, 1993, p. 124) gives an impression of Part II:

> **10.3 *Clear* Verbs**
> **Class Members:** clear, clean, drain, empty
> **Properties:**
>
> (87) Locative Alternation (transitive):
> a. Doug cleared dishes from the table. (locative variant)
> b. Doug cleared the table of dishes. (*of* variant)
>
> (90) Causative/Inchoative Alternation (except *clean*):
> a. The strong winds cleared the skies.
> The skies cleared.
>
> (91) * Resultative Phrase:
> * Doug cleared the table clean.

3 Extending LaVA

3.1 Dataset

Using LaVA (Kann et al., 2018, 2019) as a starting point, we added three new columns: one for the selected NP resultative (Res_Selected_NP), one for the non-selected NP resultative (Res_Non-Selected_NP), and—to speed up the process a bit—we created an additional column indicating whether a verb can participate in a resultative at all (Resultative).

The column Origin indicates whether the verb was already present in LaVA, or that we added it. The Res_Selected_ExampleSentence (RS) column is where Selected NP resultative sentences are entered, while non-selected NP resultative sentences go in Res_Non-Selected_ExampleSentence (RNS), along with their source in either Source RS or Source RNS. The Comments column (not shown in Table 3.1) enables us to look up certain statistics or to query all the verbs with a specific property.

VerbID	Verb	Origin	Resultative	Res_Selected_NP	Res_Non-Selected_NP	Res_VerbSource	Res_Selected_ExampleSentence (RS)	Source RS	Res_Non-Selected_ExampleSentence (RNS)	Source RNS
37	read	LaVA verb	1	x	1	L17			He read himself awake each morning	L17
102	broke	LaVA verb	1	1	x	L93-45.1	Tony broke the piggy bank open	L93-45.1		
118	burned	LaVA verb	1	1	x	L93-45.4	Amanda burned the stove black	self		
316	poured	LaVA verb	1	x	1	L17			I poured the cup full	L17
516	advanced	new verb	0	0	0	L93-51.1				
570	crowed	new verb	1	0	1	L93-38			The rooster crowed everyone awake	L93-18

TABLE 3.1: Sample of the new columns in LaVA.

3.2 New verbs

The verbs in the dataset were chosen such that each would participate in at least one frame of one alternation (Kann et al., 2019). Comprising 516 verbs in total, each frame thus has more than 100 verbs with either a positive or negative label. In order for the resultative to be as representative as the other five alternations, the number of verbs must increase.

The most intuitive place to start looking for verbs that participate in the resultative construction is Levin (1993) section 7.5 on the resultative construction. Unlike other sections, which list numerous verb classes that are associated with that particular alternation or frame, this section only lists two: Verbs of Inherently Directed Motion (51.1), and Bring and Take (11.3). Nevertheless, we take all the verbs that are not in LaVA yet.

Of the 191 verb classes in Part II, 26 are said to have the property 'Resultative Phrase,' each with one to three example sentences. We add all the verbs used in the examples, along with a small number of other verbs from that class. The verbs without an example were often selected if it was obvious to us that they would indeed be able to participate in the construction.

The original LaVA dataset consists of 516 verbs. We added 70 verbs to that list, bringing the total number to 586 verbs. Table 3.2 shows all new verbs for reference.

Verb	Section	Verb	Section	Verb	Section	Verb	Section	Verb	Section
smelled	7.5	spanked	18.3	taped	22.4	murdered	42.1	escaped	51.1
cleared	10.3	flogged	18.3	ripped	23.2	killed	42.1	exited	51.1
licked	10.4.1	thrashed	18.3	slipped	23.2	strangled	42.2	fell	51.1
wiped	10.4.1	caressed	20	colored	24	tore	45.1	fled	51.1
took	11.3	grazed	20	painted	24	boiled	45.3	plunged	51.1
drew	12	kissed	20	bored	31.1	baked	45.3	receded	51.1
shoved	12	nudged	20	crowed	38	fidgeted	49	rose	51.1
tugged	12	patted	20	breathed	40.1.2	advanced	51.1	tumbled	51.1
yanked	12	pinched	20	cried	40.2	arrived	51.1	went	51.1
knocked	18.1	prodded	20	laughed	40.2	ascended	51.1	slid	51.3.1
hit	18.1	stroked	20	coughed	40.2	came	51.1	walked	51.3.2
clubbed	18.3	stung	20	asphyxiated	40.7	departed	51.1	skated	51.4.1
knifed	18.3	tickled	20	drowned	40.7	descended	51.1	rowed	51.4.2
pummeled	18.3	touched	20	suffocated	40.7	entered	51.1	waltzed	51.5

TABLE 3.2: Verbs we added to LaVA, sorted by the Levin (1993) section the verb appears in. The verbs in the dataset are in the past tense, hence the representation here. For consistency with the existing verbs, we have maintained the use of US English spelling.

3.2.1 New verbs, existing alternations

With the 70 new verbs added to the dataset, the columns for the existing 5 alternations are not complete anymore. The new verbs must be labelled for the existing five alternations, too. We do this in the same way as Kann et al. (2019), using Levin (1993) as our only source of information pertaining to the verbs' ability to participate in the frames.

Column	1	0	x
sl	0	2	68
sl_noloc	0	2	68
sl_nowith	2	0	68
inch	10	32	28
non_inch	31	10	29
there	12	7	51
non_there	2	13	55
dat_both	6	64	0
dative_to	1	69	0
dat_do	0	70	0
refl_op	0	70	0
refl_only	0	70	0
Resultative	42	28	0
Res_Selected_NP	30	36	4
Res_Non-Selected_NP	12	29	29

TABLE 3.3: Number of label occurrences per column for the 70 new verbs.

3.3 Method

In this section, we describe how we determined which verbs can and cannot participate in the resultative construction.

3.3.1 Label requirements

Selected NP resultatives

A verb gets assigned the label '1' in this column iff that verb can be used as the main verb in:

(a) a semantically valid resultative sentence, and

(b) sentence (a) without the result. The intended semantics need not be the same.

It gets assigned the label '0' iff criterion (a) is not met.

Non-selected NP resultatives

A verb gets assigned the label '1' in this column iff:

(a) a semantically valid resultative sentence can be constructed using the verb as the main verb, and

(b) no longer forms a semantically valid sentence when the result is removed from sentence (a).

It gets assigned the label '0' iff condition (b) is not met.

In both columns, the label 'x' is given otherwise. This occurs when it is unknown whether a verb meets any of the above criteria.

We will only consider results that can be used independently to describe the current state. One could verify this by rephrasing the sentence, for instance from "I wiped the table clean" to "The table is clean" (`NP1 V NP2 Res` → `NP2 V-be Res`). This ensures that phrasal verbs like *brighten up*, *cool down*, and those with 'off' are discarded as valid options. However, 'off' can be used as part of a result when it is used physically. Also, it does not matter if a verb can have different meanings, as long as it provides a valid resultative sentence.

3.3.2 Finding the appropriate labels

First, we look only at the general Resultative column, to filter out all the verbs that, according to the literature, definitely do not take a resultative, so we can ignore those verbs in subsequent passes. This is done as follows.

1. We use Levin (1993) as a reference to get started. Section 7.5 lists arguments that resultatives can or cannot be predicated of, with examples for each of the arguments. Two verb classes are named explicitly: verbs of inherently directed motion, and bring and take. These two are unable to form a resultative sentence and are described in Part II of the book, so these verbs can easily be labelled '0'.

2. Stative verbs are also excluded from this construction (Levin, 1993), but it is not listed as a class, so we obtained three lists of stative verbs as a reference (*Stative verbs* 2020; *Stative verbs* 2014; *Stative verbs* 2022). Seventeen of the combined 74

unique verbs are present in our dataset. Since they do not occur in the resultative construction, they are labelled '0'.

3. For each verb class in Part II, its members (the verbs) and properties (the alternations) are listed along with example sentences (see the example in section 2.3). The property of 'Resultative Phrase' is given under 33 verb classes (26 positive and 7 negative), which we use to label the verbs listed in 'Class Members' accordingly.

   Sometimes, the book notes that the resultative is a property of "some verbs" or "most verbs". This means we cannot label all the listed verbs. In those cases, we label the uncertain verbs with 'x' until a resultative phrase confirming its participation is found. We also document this in the Comments column. (This applies to 13 of 568 verbs in the dataset.)

4. Having exhausted the 1993 book, we turn to two of Levin's conference papers (Levin, 2015, 2017) on resultatives. These documents contain additional information on some verbs. The approach of manually searching and labelling the listed verbs is like in step 3. Other resources we use are Boas (2003) and Goldberg and Jackendoff (2004).

5. Whenever conflicting information is encountered, the positive claim takes precedence over the negative, since only one positive example is needed to prove that a resultative sentence can be formed. For instance, according to Levin (1993, p. 155) section 20, *kiss* cannot take a result as argument, but Boas (2003, p. 15) shows that it can.

6. All remaining verbs are labelled 'x'. These will be revisited at the end of the process.

Now that the Resultative column is filled, we can look closer at the selected and non-selected NP resultatives. Because we have the information about a verb's ability to form a resultative sentence from the steps above, we can ignore the '0' and 'x' labels and consider only the verbs labelled '1'. For each "positive" verb, we tried to find a sentence that proves that said verb can, indeed, participate in a (non-)selected NP resultative frame.

7. For each verb, we search the same sources as before again for sentences containing that particular verb. If such a sentence does not appear in those sources, we search through the "Example sentences" section in the verb's Cambridge Dictionary entry (*Cambridge Dictionary* 2024). The sentence, its source, and the label are then documented in their appropriate locations.

8. Some verbs do not appear in a resultative context, but may have an obvious example sentence. These sentences are entered with the source being "self".

Once all verbs with Resultative = '1' are labelled, we return to the verbs we labelled 'x' in step 6. We go through these verbs one final time to see if we could come up with sentences that show that they can take a resultative. When all resources have been exhausted, all remaining verbs get the label 'x' in Res_Selected_NP and Res_Non-Selected_NP.

3.3.3 Exceptions

Adjectives as verbs

Many ‘change of state’ verbs are zero-related to adjectives (Levin, 1993, p. 28), which makes it difficult to find sentences that feel natural. Section 45.4 of Levin (1993, p. 245) claims that these verbs can participate in a resultative sentence. Fourteen of the 35 verbs in section 45.4 are included in LaVA. We could find sentences for only three of those.

Double-word entries

LaVA includes six entries that consist of two words. Four of these appear of the form `Verb Result` and these are paired with an entry that only contains the verb. Table 3.4 below shows all ten of the verbs with their combinations. We have decided to label all the verbs on the left as negative in all three resultative columns (0, 0, 0), because those verbs already have an argument. It is not possible to take the result as a second argument, e.g. “* I covered up the hole up” or “* I covered up up the hole.”

VerbID	Verb	VerbID	Verb
101	flung open	33	flung
107	tipped over	78	tipped
208	covered up	250	covered
288	stopped up	113	stopped
409	flunked out		
423	had faith		

TABLE 3.4

3.4 Result

Ultimately, close to half of the verbs (269 of 586) have been assigned a definitive label (i.e. 0 or 1) in the Resultative column. This percentage is comparable with some of the existing five alterations in the dataset. The Selected frame ends up with 204, while Non-Selected has the smallest number of labels of all the frames, totalling just 110 labels. Table 3.5 shows the distribution of positive and negative labels for all verb frames in LaVA.

Alternation	CAUSATIVE-INCHOATIVE		DATIVE		SPRAY-LOAD		*there*-INSERTION		UNDERSTOOD-OBJECT		RESULTATIVE CONSTRUCTION		
	Inchoative	Causative	Preposition	Double-Object	With	Locative	No-There	There	Reflexive	Non-Reflexive	Resultative	Selected	Non-Selected
positive	83	165	72	80	101	88	164	62	84	11	210	129	50
negative	176	0	440	506	244	257	0	199	489	573	59	75	60
Total	259	165	512	586	345	345	164	261	573	584	269	204	110

TABLE 3.5: Verb membership class distributions for each frame after executing 3.2 and 3.3.

3.4.1 Note on the number of labels

Figure 3.1 shows that the number of 0’s, 1’s and x’s varies considerably. The frames in the DATIVE and UNDERSTOOD-OBJECT alternations have a lot of ‘0’ labels (ca. 500), whereas the number of 0’s and 1’s in the Resultative frames are more equal to the CAUSATIVE-INCHOATIVE and *there*-INSERTION alternations (ca. 200). SPRAY-LOAD sits in between (ca. 350). A truth value is necessary for training and testing a classifier. Since the ‘x’ values

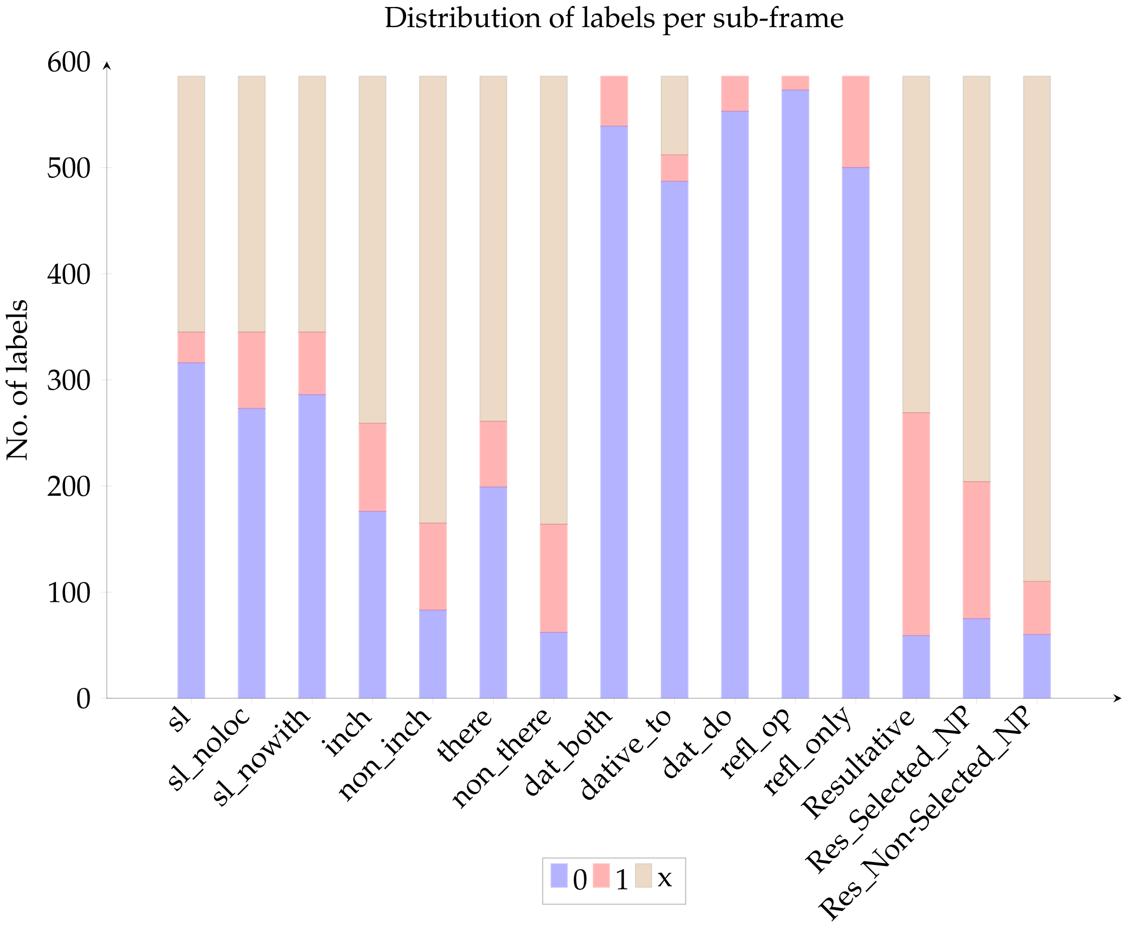


FIGURE 3.1: Distribution of the labels in each column of LaVA. This graphical representation shows the differences between the alternation well.

are defined as unknown, the verbs with a label 'x' in that specific frame are removed before training. This may have major consequences for the performance.

4 Experiment background

We want to run a simple experiment to measure how well the augmented dataset holds up against the original one. Perhaps we are even able to make some conclusive claims about the resultative construction in verb embeddings. Before we can describe the experiment we carried out, we need to decide on the design. We will first introduce the two experiments it is based on in section 4.1 and section 4.2, followed by a brief summary of the design considerations (section 4.3).

4.1 Kann et al. (2019) Experiment 1

This paper describes two experiments, but only the first one is relevant for our experiment. In that experiment, they aim to classify (un)acceptable verb–frame combinations based solely on the word embeddings.

Two types of word embeddings are used in this experiment:

- pre-trained 300-dimensional GloVe word embeddings trained on 6B tokens (Pennington et al., 2014);
- word embeddings trained by an LSTM (Warstadt et al., 2019) on the 100M token British National Corpus (BNC; Leech, 1993). The LSTM learns word embeddings for the 100k most frequent words in the BNC. These embeddings are referred to as "CoLA-style" embeddings, after the project they were trained for (Kann et al., 2019; Warstadt et al., 2019, 2018).

The experiment is essentially performed twice, once with CoLA-style embeddings and once with GloVe embeddings.

Experiment To attain their goal, Kann et al. train one classifier per alternation, where the frames in Table 1.1 are the classes to predict. Thus, each classifier predicts (non-)membership for the two different classes belonging to an alternation. A multi-layer perceptron (MLP) with a single hidden layer and 30-dimensional hidden states is used to model the probability of a frame being acceptable for a given verb. The Adam optimizer (Kingma and Ba, 2015) is used during training and all ANNs are trained for 15 epochs. Of all the verbs, one half is used as training data, while the other half is split equally into a development set and a test set, using 4-fold cross-validation.

Evaluation The models are evaluated using accuracy and Matthews correlation coefficient (MCC; Matthews, 1975), which gives a value between -1 and 1 for the correlation between two binary distributions. A value of 1 means a perfect correlation, whereas -1 represents an inverse prediction. A score of 0 indicates no relationship between two distributions. MCC's disregard for class size differences makes it more robust to classification of unbalanced distributions than F1-score or accuracy, which favour classifiers with a majority class bias.

	CoLA			GloVe		
Frame	Accuracy	Majority baseline	MCC	Accuracy	Majority baseline	MCC
Inchoative	0.810	0.667	0.555	0.855	0.668	0.672
Causative	1.000	1.000	0.000	1.000	1.000	0.000
Preposition	0.866	0.850	0.320	0.850	0.850	0.000
Double-Object	0.883	0.849	0.482	0.853	0.853	0.000
With	0.858	0.710	0.645	0.893	0.710	0.585
Locative	0.729	0.739	0.253	0.734	0.746	0.145
No-There	1.000	1.000	0.000	1.000	1.000	0.000
There	0.843	0.787	0.459	0.858	0.791	0.536
Reflexive	0.977	0.977	0.000	0.976	0.976	0.000
Non-Reflexive	0.790	0.830	0.219	0.732	0.815	0.300
Resultative	-	-	-	-	-	-
Selected	-	-	-	-	-	-
Non-Selected	-	-	-	-	-	-

TABLE 4.1: Results from Kann et al. (2019) for the CoLA and GloVe embeddings. The bottom three frames were not present in the original paper, but are included here for consistency with subsequent tables.

They also established a majority baseline, where the most frequent class is predicted for every verb. This is a sanity check to assert that the accuracy has improved.

Results The accuracies for the GloVe and CoLA-style embeddings are comparable for all classes, suggesting that they contain similar information about verbs and syntactic frames. According to Kann et al., this would make sense, as both embeddings are based on co-occurrences of words.

The GloVe embeddings for the Causative, There and both DATIVE frames have an MCC of 0, meaning that the model predictions are about as good as random. All other classes obtain a weak (0.1–0.5) to moderate (0.5–0.7) MCC, indicating that information about the syntactic frames can be extracted from verb embeddings. Relatively good performance (> 0.45) is found for the Inchoative, With and No-There frames for both CoLA and GloVe embeddings, as well as Double-Object using the CoLA embeddings.

4.2 Yi et al. (2022) Experiment 1

The second paper related to our experiment investigates the extent to which word and sentence embeddings encode verb alternation classes. The authors hypothesise that the alternations should be observable within large text corpora, and can therefore be used in the pre-training of Pre-trained Language Models (PLMs). Just as in section 4.1, only the first of the two experiments is relevant for our purposes.

Whereas Kann et al. only use static embeddings, Yi et al. (2022b) make use of both static and contextual embeddings, though the contextual embeddings are very much the main focus.

BERT is trained on the BooksCorpus (800M words) (Zhu et al., 2015) and the English Wikipedia (2500M words). BERT-base has 12 layers and 12 attention heads, and its input embeddings have 768 dimensions (Devlin et al., 2019). In the token-embedding layer (layer 0), the verb embeddings are simply the pre-trained input token embeddings that correspond to each verb. For layers 1–12, contextual information from the sentences in FAVA is included in the verb embedding. The authors are specifically interested in the differences between the layers, so sentences in FAVA containing the verb are used to also form a "layer-embedding" for each verb.

Experiment The experiments are performed on several Transformer-based PLMs: BERT, DeBERTa, ELECTRA, and RoBERTa. Each model varies in tokenisation, pre-training and (the size of) their training corpus. The base architectures for each model are used to make comparisons between them fair. These architectures all have 12 layers, 12 attention heads and a hidden layers size of 768.

To find out if PLM word embeddings encode information about which frames a verb can participate in, the researchers employ a logistic regression classifier for each syntactic frame. These classifiers take a verb's layer embedding representation as input and predict whether the verb can participate in their respective frames (e.g. for the SPRAY-LOAD alternation, one classifier for the Locative frame and one for the With frame).

Following Kann et al. (2019), the data is split into four equally-sized folds, but instead of spending a quarter of the verbs on a development set, Yi et al. (2022b) chose to use three folds for the training set, and one for the test set.

Frame	Accuracy	Majority baseline	MCC
Inchoative	0.885	0.664	0.741
Causative	1.000	1.000	0.000
Preposition	0.925	0.853	0.701
Double-Object	0.913	0.857	0.614
With	0.848	0.706	0.633
Locative	0.834	0.749	0.525
No-There	1.000	1.000	0.000
There	0.876	0.793	0.593
Reflexive	0.867	0.833	0.466
Non-Reflexive	0.984	0.979	0.563
Resultative	-	-	-
Selected	-	-	-
Non-Selected	-	-	-

TABLE 4.2: Result from Yi et al. (2022a) for the static embeddings. The bottom three frames were not present in the original paper, but are included here for consistency with subsequent tables.

Results Overall, they find that both the MCC and accuracy of the contextual PLM embeddings are greatly improved compared to the reference CoLA-style embeddings. The PLMs perform well even on the—for the CoLA embeddings—more difficult frames, with BERT achieving 0.969 MCC on the Locative frame (CoLA embedding: 0.253 MCC). The different layers of each PLM show consistent patterns in performance.

However, these results are based on the best layer using contextual embeddings.[3] Yi et al. (2022b, Figures 1 and 2) show that the static layer is by far the worst one in terms of MCC. Table 4.2 shows the actual results for the static embeddings.

[3] The caption of Yi et al. (2022b) Table 2 contains an error: these are the results from Word-Level experiments with *contextual* embeddings, not with *static* embeddings.

4.3 Considerations

At first glance, Kann et al. (2019)'s Experiment 1 (4.1) looks to be the simplest experiment, since it only involves static word embeddings and a multi-layer perceptron with one hidden layer. A downside of this experiment is that it is not immediately clear what the output would look like. Source code for the experiments in this paper is not available.

For Yi et al. (2022b)'s Experiment 1 (4.2), the source code *is* available. This makes it possible to play with and run the code, which gives an idea of what the input and output is like. Another major advantage is that we can directly run their experiment on our dataset, offering the most accurate like-for-like comparison.

Both experiments have similar goals: "identify the syntactic frames in which a verb can appear." Both experiments take word embeddings as input, but Yi et al. use contextual word embeddings based on the sentences in FAVA. Now, we do have a tiny dataset with resultative phrases (see section 3.3), but it is nowhere near as extensive as the existing FAVA. Moreover, the experiment should be as simple as possible, so this is beyond the scope of this thesis.

Ultimately, we decided to use the code (Yi et al., 2022a) written by Yi et al. (2022b) as a starting point, since it is similar to what we want to do, and it can be adjusted to our needs with relative ease. In essence, we are partly reproducing their experiment; validation of their findings could be a supplementary objective.

5 Experiment

Originally, in section 3.3.2, we created the Resultative column just to eliminate verbs that would not be able to participate in the resultative construction. This means that this frame contains information about whether a verb can form a resultative sentence *at all*. Since it will always be a superset of the Selected and Non-Selected frames, we expect that the model will perform better on the Resultative than on the other two frames.

In the experiment, each frame of the alternation is classified independently. The classifier does not actually predict if a verb can participate in an alternation, but if it can participate in a frame. Hence, there will be 13 frames in total: 10 for the existing five alternations, and 1 for each of Resultative, Res_Selected_NP and Res_Non-Selected_NP.

Before we can run the experiment, we establish a baseline, explained in section 5.1. Section 5.2 describes the actual experiment. The results are presented in the final section (5.3).

5.1 Baseline experiment

Since the dataset for resultative frame membership did not exist at the time when Yi et al. carried out their experiments, there are no direct results in the paper that we can compare with. In order to do that, we need some sort of baseline. As their code is publicly available (Yi et al., 2022a), we can run (a slightly modified version of) it to obtain an estimate of what their results would have been if they had had our dataset. This baseline enables us to make a fair and real comparison with their results.

In this baseline experiment, we run a linear classifier on a version of LaVA which includes the resultative columns, but no new verbs (i.e. only the 516 original verbs). Since each frame is predicted and evaluated individually, the results for the 10 pre-existing frames should be exactly the same as in Yi et al. (2022b). However, it is the three resultative frames we are interested in. The results from the experiment (in the next section) with our full extended dataset will be compared against the results from *this* baseline. As noted in section 2.1.1, the frames Causative and No-There have no negative examples, and therefore no results. The frames are included for completeness.

Results Table 5.1 shows the results for the baseline experiment. As expected, the results for the frames Inchoative – Non-Reflexive are exactly identical to the results in Table 4.2. The new frames score an average accuracy of 0.843 and MCC of 0.589, just below the 0.892 accuracy and 0.604 MCC for the existing frames (excluding Causative and No-There). For all frames, the accuracy is higher than the majority baseline.

Frame	Accuracy	Majority baseline	MCC
Inchoative	0.885	0.664	0.741
Causative	1.000	1.000	0.000
Preposition	0.925	0.853	0.701
Double-Object	0.913	0.857	0.614
With	0.848	0.706	0.633
Locative	0.834	0.749	0.525
No-There	1.000	1.000	0.000
There	0.876	0.793	0.593
Reflexive	0.867	0.833	0.466
Non-Reflexive	0.984	0.979	0.563
Resultative	0.905	0.844	0.615
Selected	0.841	0.717	0.591
Non-Selected	0.783	0.551	0.560

TABLE 5.1: Performance metrics for the baseline experiment, using all frames in LaVA without the new verbs.

5.2 Setup

5.2.1 Model

We choose to use the BERT base model in the experiment. The other models used in Yi et al. (2022b) (DeBERTa, ELECTRA, RoBERTa) performed very similarly, so there is no particular reason to choose either of those over BERT. We did consider using a more recent Large Language Model (LLM), e.g. LLaMa (Touvron et al., 2023), but the disk space and computational requirements exceeded what we have available. Furthermore, LLaMa does not provide access to the input embeddings, which is a key requirement for our experiment.

BERT is trained on the BooksCorpus (800M words) (Zhu et al., 2015) and the English Wikipedia (2500M words). BERT-base has 12 layers and 12 attention heads, and its input embeddings have 768 dimensions (Devlin et al., 2019). However, in this experiment we only consider the output after the first layer (layer 0). If we only take the first (token-)embedding layer, we do not need a FAVA-like dataset for context of the word embeddings, which is how the verb embeddings are created for layers 1–12 (Yi et al., 2022a,b).

5.2.2 Method

First, we retrieve the static word embeddings created by BERT. The classifier uses Logistic Regression without regularisation as implemented in scikit-learn (Pedregosa et al., 2011). We employ 4-fold cross-validation, where 3 folds are used as the training set and 1 fold for the test set. The classifier is trained on the verb frames data using the word embeddings as features.

The accuracy and MCC are calculated for each frame. As discussed in section 3.4.1, there is a sizeable class imbalance in the data. MCC is arguably the most informative metric: a high MCC means high values in other metrics as well (Chicco et al., 2021). Hence, we will use MCC as our primary method for evaluation.

5.3 Results

In Table 5.2, we present the accuracy and MCC from the experiment on the full (extended) LaVA dataset, i.e. with all 586 verbs and all 13 frames. Table 5.3 shows the difference with the results from the baseline (Table 5.1).

Frame	Accuracy	Majority baseline	MCC
Inchoative	0.884	0.680	0.732
Causative	1.000	1.000	0.000
Preposition	0.912	0.859	0.614
Double-Object	0.908	0.863	0.567
With	0.852	0.707	0.638
Locative	0.826	0.745	0.510
No-There	1.000	1.000	0.000
There	0.847	0.762	0.548
Reflexive	0.871	0.853	0.412
Non-Reflexive	0.986	0.981	0.519
Resultative	0.855	0.781	0.543
Selected	0.794	0.632	0.549
Non-Selected	0.836	0.545	0.670

TABLE 5.2: Performance metrics for all frames in LaVA including the new verbs.

Looking at the 10 original frames first, we observe that the accuracy remains relatively stable compared to the baseline, although the general trend is downwards. The MCCs on the other hand, are lower for all frames except one. Only the With frame performs better, but that increase of MCC (+0.005) is much smaller than the decreases in all the other frames (with an average of −0.042). Keeping in line with the boundaries set by Kann et al. (2019), we only obtain a weak (0.3–0.5) MCC for Reflexive. The Inchoative frame achieves a strong (> 0.7) MCC, while all other frames demonstrate a moderate (0.5–0.7) correlation.

As for the three resultative frames, the accuracy of the Resultative frame is indeed higher than that of Selected and Non-Selected, as we hypothesised in Chapter 5. Interestingly, its MCC is the lowest of the three, scoring 0.543, with Selected closely following with an MCC of 0.549. The highest correlation of the three was achieved for the Non-Selected frame (0.670).

Resultative and Selected have a lower accuracy (−0.050, −0.046) and MCC (−0.072, −0.042) than the baseline. The only frame that has reasonably improved is Non-Selected, with a remarkable +0.110 MCC and a modest +0.054 increase in accuracy. While these deltas are minuscule compared to the jump from CoLA to the baseline—which use two entirely different models—we must consider the changes within the same model.

The accuracy is higher than the majority baseline on all frames tested. Compared to the baseline, the MCC of Preposition and Resultative have decreased the most: −0.086 and −0.072, respectively. Non-Selected has increased by 0.110.

Frame	Accuracy Δ	Majority baseline Δ	MCC Δ
Inchoative	-0.001	0.016	-0.009
Causative	0.000	0.000	0.000
Preposition	-0.013	0.006	-0.086
Double-Object	-0.005	0.007	-0.047
With	0.004	0.002	0.005
Locative	-0.008	-0.004	-0.015
No-There	0.000	0.000	0.000
There	-0.029	-0.031	-0.045
Reflexive	0.004	0.020	-0.054
Non-Reflexive	0.002	0.003	-0.044
Resultative	-0.050	-0.064	-0.072
Selected	-0.046	-0.085	-0.042
Non-Selected	0.054	-0.005	0.110

TABLE 5.3: Difference between the baseline and the experiment. It may seem that there are some arithmetic/rounding errors, that is because the difference is calculated using 6 decimals, while only 3 are shown here. (The raw values can be found in Appendix B.)

6 Discussion

In this chapter, we will discuss the results from both the new dataset (section 3.4) and the experiment (section 5.3). We begin by highlighting some observations in section 6.1. The limitations will be contemplated in section 6.2. We end this chapter with suggestions for future work (section 6.3).

6.1 Interpretations of the results

The performance metrics suggest that the new verbs in the dataset do not particularly benefit the performance of the model. All frames except two have worse MCC with the extra verbs than without. A possible reason is that we have mislabelled the new verbs in section 3.2. A test run of the experiment at an earlier stage supports this. In section 3.3.3, we argued that the six 'verbs' with two words should be labelled (0, 0, 0). The test run shows that, with the labels (x, x, x), the accuracy is 0.08 higher for the Resultative and Selected frames, and the MCC increase varies from 0.06 to 0.09. However, for Non-Selected, the accuracy and MCC barely change.

Only Non-Selected has increased, while the others decreased. The confusion matrices in Figure 6.1 show that for Resultative and Selected, by far the most frequent true label is '1', which is also predicted most often. This is true for both the baseline and the experiment. For the Non-Selected frame, however, the true positives and true negatives are much closer in number. In the experiment, the true negatives even surpasses the true positives.

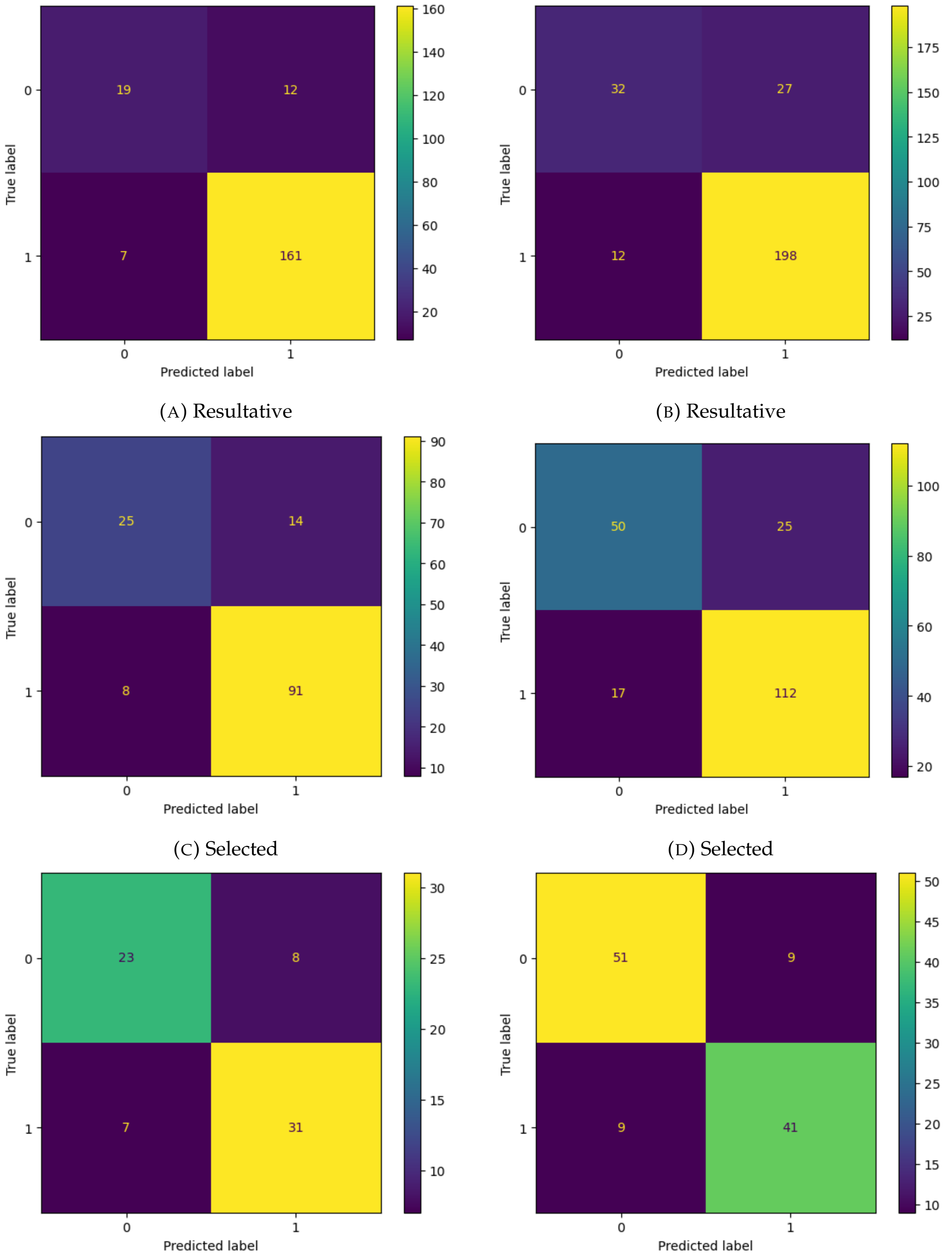
As the total number of verbs grows from the baseline to the experiment, so do the number of false positives and false negatives in the Resultative and Selected frames, but not for the Non-Selected frame. The change of true positive, true negative, false positive, false negative from the baseline to the experiment perfectly reflects the change in MCC. Using the formula

$$\frac{(TP+TN)_{\text{experiment}} - (TP+TN)_{\text{baseline}}}{(FP+FN)_{\text{experiment}} - (FP+FN)_{\text{baseline}}}$$

we obtain the following.

- Resultative: $\frac{(198+32)-(161+19)}{(27+12)-(12+7)} = \frac{50}{20} = 2.5$
- Selected: $\frac{(112+50)-(91+25)}{(25+17)-(14+8)} = \frac{46}{20} = 2.3$
- Non-Selected: $\frac{(41+51)-(31+23)}{(9+9)-(8+7)} = \frac{38}{3} = 12.\overline{6}$

The ratio between the correct predictions and incorrect predictions is clearly better for the Non-Selected frame.



(A) Resultative

(B) Resultative

(C) Selected

(D) Selected

(E) Non-Selected

(F) Non-Selected

FIGURE 6.1: Confusion matrices for the baseline (left) and the experiment (right).

6.2 Limitations

Dataset sparsity As Kann et al. (2019) already addressed, Levin (1993) is not a comprehensive database for all verbs and all alternations. An example sentence is given for only a subset of the verbs that participate in an alternation. In some cases, the addendum "some verbs" or "most verbs" is given to a certain alternation, leaving the reader guessing which verbs do and do not participate. Since the LaVA dataset primarily relies on information in Levin (1993), not all verbs have a positive or negative judgement. Despite consulting other sources in addition to Levin (1993), the resulting dataset is still sparse (see section 3.4).

Embeddings Yi et al. (2022b) clearly shows that PLM contextual embeddings yield significantly better results than static embeddings. The experiment presented in Chapter 5 is deliberately simple, intended to get an impression of the dataset's functioning. Due to personal factors, time constraints and feasibility, we decided not to pursue the creation of a resultative sentence corpus, in favour of the lexical dataset. The context of the sentence corpus would have produced better embeddings, leading to potentially better results.

Reliability Labelling the verbs for participation in any of the resultative frames was done entirely manually. Although the majority of those labels are supported by sentences collected from literature or corpora, 11 verbs are judged by the author, who is not a native English speaker. Additionally, 'self' is credited for 50 out of 192 sentences in LaVA (see section 3.3). Though the sentences are not the intended purpose of the dataset, they do support the claims for frame participation.

However, these limitations do not detract from the validity of the study. The creation of a dataset for selected NP and non-selected NP resultatives was an essential element in our investigation of resultatives in verb embeddings. With regards to the sparsity, all previous works that depend either on Levin (1993) or Kann et al. (2019) faced the same limitation.

6.3 Future work

In this thesis, we have incremented the number of alternations in LaVA by one. This is still just a fraction of the number of *documented* alternations in English. It would be valuable to see if our findings coincide with other alternations currently not included in LaVA. As of yet, the dataset contains frame information for 586 verbs. Levin (1993) provides verb class properties for over 3100 verbs. In our research, we found that having more verbs is not necessarily advantageous. A future study could build upon this by evaluating more verbs.

From the limitations, one line of future work evidently emerges. It would be interesting to build a synthetic corpus for resultative sentences in a similar fashion as the FAVA sentence dataset was created, i.e. by generating sentences from lexical sets (see section 2.1.2). One could use the example sentences in LaVA as a starting point and simply mix-and-match the various parts-of-speech available.

Finally, as we alluded to in section 5.2.1, we briefly considered using LLaMa for the model. With the rise in popularity of Large Language Models (LLMs) (Ignat et al., 2023; Zhao et al., 2023)—in no small part due to their exceptional performance (Liu et al., 2023; OpenAI et al., 2024)—investigating the identification of verb alternations using an LLM would be interesting.

7 Conclusion

In this thesis we investigate if the resultative construction—specifically selected NP resultatives and non-selected NP resultatives—can be represented in PLM word embeddings. We present an extended version of the lexical dataset LaVA, which now includes data on 586 verbs and their participation in the two aforementioned types of resultatives. We train a logistic regression classifier on the verb embeddings to distinguish each verb–frame combination. As previous research showed that verb embeddings could be used to predict other alternations, we expected that the resultative construction would not be any different, even though less data for it was available. We find that the performance metrics are on par with the existing alternations, albeit slightly worse when the new verbs are included.

As introduced in section 1.1, LaVA and FAVA covered only five alternations. With this thesis, the word-level dataset (LaVA) is one alternation richer. The achievements from this thesis can be built upon in the future, bolstering the knowledge and understanding of verb frames and their representations in language models.

Bibliography

Alexiadou, Artemis and Florian Schäfer (2011). "There-Insertion: An Unaccusativity Mismatch at the Syntax-Semantics Interface". In: *Online proceedings of WCCFL* 28. URL: https://www.researchgate.net/publication/268424565_There-Insertion_An_Unaccusativity_Mismatch_at_the_Syntax-Semantics_Interface.

Arad, Maya (2006). "The Spray-Load Alternation". In: *The Blackwell Companion to Syntax*. Section: 63. John Wiley & Sons, Ltd, pp. 466–478. ISBN: 978-0-470-99659-1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470996591.ch63 (visited on 2024-07-12).

Boas, Hans C. (Feb. 1, 2003). "Appendix A - Resultative Phrases". In: *A Constructional Approach to Resultatives*. Stanford Monographs in Linguistics. University of Chicago Press. ISBN: 978-1-57586-408-2. URL: https://web.stanford.edu/group/cslipublications/cslipublications/hand/1575864088appendix.pdf (visited on 2023-11-21).

Cambridge Dictionary (June 12, 2024). *Cambridge Dictionary | English Dictionary, Translations & Thesaurus*. URL: https://dictionary.cambridge.org/ (visited on 2024-06-30).

Chicco, Davide, Niklas Tötsch, and Giuseppe Jurman (Feb. 4, 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". In: *BioData Mining* 14.1, p. 13. ISSN: 1756-0381. DOI: 10.1186/s13040-021-00244-z. URL: https://doi.org/10.1186/s13040-021-00244-z (visited on 2024-07-12).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (visited on 2024-04-24).

Dikken, Marcel den and Eric Hoekstra (Jan. 1, 1994). "No cause for a Small Clause? (Non-)arguments for the structure of resultatives". In: *GAGL : Groninger Arbeiten zur germanistischen Linguistik* 37, pp. 89–105. ISSN: 0924-655X. URL: https://ugp.rug.nl/GAGL/article/view/30309 (visited on 2024-05-24).

Fillmore, Charles J. (1970). "The Grammar of Hitting and Breaking". In: *Readings in English Transformational Grammar*. Ed. by Roderick A. Jacobs and Peter S. Rosenbaum. Publisher: Waltham, Massachussets, pp. 120–133. URL: https://www1.icsi.berkeley.edu/pubs/ai/ICSI_grammarofhitting12.pdf (visited on 2023-07-20).

Gildea, Daniel and Daniel Jurafsky (2002). "Automatic Labeling of Semantic Roles". In: *Computational Linguistics* 28.3, pp. 245–288. URL: http://www.cs.rochester.edu/~gildea/gildea-cl02.pdf.

Goldberg, Adele E. and Ray Jackendoff (2004). "The English Resultative as a Family of Constructions". In: *Language* 80.3. Publisher: Linguistic Society of America, pp. 532–568. ISSN: 0097-8507. URL: https://www.jstor.org/stable/4489722 (visited on 2023-04-14).

Hugging Face (Oct. 3, 2022). *BERT base model (uncased)*. URL: https://huggingface.co/google-bert/bert-base-uncased (visited on 2024-07-01).

Ignat, Oana, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea (May 21, 2023). *A PhD Student's Perspective on Research in NLP in the Era of Very Large Language Models*. DOI: 10.48550/arXiv.2305.12544. arXiv: 2305.12544[cs]. URL: http://arxiv.org/abs/2305.12544 (visited on 2023-05-26).

Jurafsky, Daniel and James H. Martin (Feb. 3, 2024a). "Semantic Role Labeling". In: *Speech and Language Processing*. 3 (draft). Chapter 20. URL: https://web.stanford.edu/~jurafsky/slp3/20.pdf (visited on 2024-06-27).

– (Feb. 3, 2024b). "Transformers and Large Language Models". In: *Speech and Language Processing*. 3 (draft). Chapter 10. URL: https://web.stanford.edu/~jurafsky/slp3/10.pdf (visited on 2024-04-12).

Kann, Katharina, Alex Warstadt, Adina Williams, and Samuel R. Bowman (2018). *Lexical Verb–Frame Alternations dataset*. URL: https://nyu-mll.github.io/CoLA/lava.zip (visited on 2024-06-30).

– (2019). "Verb Argument Structure Alternations in Word and Sentence Embeddings". In: *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*. SCiL 2019, pp. 287–297. DOI: 10.7275/q5js-4y86. URL: https://aclanthology.org/W19-0129 (visited on 2023-01-23).

Kilgarriff, Adam (Mar. 15, 1996). *BNC database and word frequency lists*. URL: https://www.kilgarriff.co.uk/bnc-readme.html (visited on 2024-07-07).

Kingma, Diederik P. and L. Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: Publisher: Ithaca, NYArXiv. URL: https://dare.uva.nl/search?identifier=a20791d3-1aff-464a-8544-268383c33a75 (visited on 2024-11-13).

Lawler, John M. (n.d.). *Reverse-engineered list of Levin Categories*. URL: https://websites.umich.edu/~jlawler/levin.verbs (visited on 2024-07-02).

Leech, Geoffrey (Jan. 1993). "100 million words of English". In: *English Today* 9.1, pp. 9–15. ISSN: 1474-0567, 0266-0784. DOI: 10.1017/S0266078400006854. URL: https://www.cambridge.org/core/journals/english-today/article/abs/100-million-words-of-english/8CD684387860612E6187FA41A99BAA01 (visited on 2024-07-07).

Levin, Beth (Sept. 1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press. 366 pp. ISBN: 978-0-226-47533-2. URL: https://press.uchicago.edu/ucp/books/book/chicago/E/bo3684144.html (visited on 2023-05-02).

– (Apr. 17, 2015). "The ingredients of nonselected NP resultatives". In: *University of Utah Student Conference of Linguistics (UUSCIL) 2017*. UUSCIL 2017. University of Utah, Salt Lake City, UT. URL: http://web.stanford.edu/~bclevin/utah15res.pdf (visited on 2023-04-14).

– (June 29, 2017). "Resultatives and Causatives". In: *Linguistic Perspectives on Causation Workshop*. The Hebrew University of Jerusalem, Jerusalem, Israel. URL: http://web.stanford.edu/~bclevin/jer17res.pdf (visited on 2023-04-14).

Levin, Beth and Malka Rappaport Hovav (1995). *Unaccusativity: at the syntax-lexical semantics interface*. Linguistic inquiry monographs 26. Cambridge, Massachussets: The MIT Press. xiii, 336. ISBN: 978-0-262-62094-9. URL: https://mitpress.mit.edu/9780262620949/unaccusativity/.

Liu, Xiao, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang (Aug. 26, 2023). "GPT understands, too". In: *AI Open*. ISSN: 2666-6510. DOI: 10.1016/

j.aiopen.2023.08.012. URL: https://www.sciencedirect.com/science/article/pii/S2666651023000141 (visited on 2024-07-14).

Loáiciga, Sharid, Luca Bevacqua, and Christian Hardmeier (Nov. 1, 2021). "Unsupervised Discovery of Unaccusative and Unergative Verbs". In: *ArXiv*. URL: https://www.semanticscholar.org/paper/Unsupervised-Discovery-of-Unaccusative-and-Verbs-Lo'aiciga-Bevacqua/b2cf51d9461d7409556df971e3eff2ff9a0d96a4 (visited on 2023-01-18).

Matthews, Brian W. (Oct. 20, 1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2, pp. 442–451. ISSN: 0005-2795. DOI: 10.1016/0005-2795(75)90109-9. URL: https://www.sciencedirect.com/science/article/pii/0005279575901099 (visited on 2024-07-02).

matthews_corrcoef (n.d.). scikit-learn. URL: https://scikit-learn/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html (visited on 2024-07-05).

National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Science, Engineering, Medicine, and Public Policy, Board on Research Data and Information, Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Analytics, Division on Earth and Life Studies, Nuclear and Radiation Studies Board, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Board on Behavioral, Cognitive, and Sensory Sciences, and Committee on Reproducibility and Replicability in Science (May 7, 2019). "Replicability". In: *Reproducibility and Replicability in Science*. National Academies Press (US). URL: https://www.ncbi.nlm.nih.gov/books/NBK547524/ (visited on 2024-11-12).

OpenAI et al. (Mar. 4, 2024). *GPT-4 Technical Report*. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774[cs]. URL: http://arxiv.org/abs/2303.08774 (visited on 2024-07-14).

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 2024-07-10).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: http://aclweb.org/anthology/D14-1162 (visited on 2024-04-12).

Piñón, Christopher (Oct. 3, 2001). "A Finer Look at the Causative-Inchoative Alternation". In: *Semantics and Linguistic Theory* 11. ISSN: 2163-5951. DOI: 10.3765/salt.v11i0.2858. URL: http://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/2858 (visited on 2023-04-21).

Richards, Norvin (2017). *Resultatives*. Part of MIT course 24.902. URL: https://web.mit.edu/norvin/www/24.902/resultatives.html (visited on 2024-01-16).

Schäfer, Florian (Mar. 2009). "The Causative Alternation". In: *Language and Linguistics Compass* 3.2, pp. 641–681. ISSN: 1749-818X, 1749-818X. DOI: 10.1111/j.1749-818X.2009.00127.x. URL: https://compass.onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2009.00127.x (visited on 2024-07-12).

Seyffarth, Esther (2019). "Identifying Participation of Individual Verbs or VerbNet Classes in the Causative Alternation". In: *Proceedings of the Society for Computation in Linguistics*

(SCiL) 2019. SCiL 2019. Ed. by Gaja Jarosz, Max Nelson, Brendan O'Connor, and Joe Pater, pp. 146–155. DOI: 10.7275/efvz-jy59. URL: https://aclanthology.org/W19-0115 (visited on 2024-07-14).

Seyffarth, Esther and Laura Kallmeyer (Dec. 2020). "Corpus-based Identification of Verbs Participating in Verb Alternations Using Classification and Manual Annotation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. COLING 2020. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4044–4055. DOI: 10.18653/v1/2020.coling-main.357. URL: https://aclanthology.org/2020.coling-main.357 (visited on 2023-01-31).

Simpson, Jane (Jan. 1, 1983). "Resultatives". In: *Papers in Lexical-functional Grammar*. Ed. by Lori Levin, Malka Rappaport Hovav, and Annie Zaenen. Accepted: 2005-10-24. Indiana University Linguistics Club, pp. 143–157. URL: https://ses.library.usyd.edu.au/handle/2123/140 (visited on 2023-05-05).

Stative verbs (Mar. 29, 2020). British Council | LearnEnglish. URL: https://learnenglish.britishcouncil.org/grammar/b1-b2-grammar/stative-verbs (visited on 2023-05-08).

Stative verbs (Mar. 3, 2014). *Stative Verbs - List of Stative Verbs & Exercises*. Ginger Grammar Rules. URL: https://www.gingersoftware.com/content/grammar-rules/verbs/stative-verbs/ (visited on 2023-05-08).

Stative verbs (Aug. 4, 2022). *Stative Verbs | List And Example Sentences*. Games4esl. Section: Vocabulary. URL: https://games4esl.com/stative-verbs-list-and-examples/ (visited on 2023-05-08).

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample (Feb. 27, 2023). *LLaMA: Open and Efficient Foundation Language Models*. DOI: 10.48550/arXiv.2302.13971. arXiv: 2302.13971[cs]. URL: http://arxiv.org/abs/2302.13971 (visited on 2024-07-10).

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman (2019). "Neural Network Acceptability Judgments". In: *Transactions of the Association for Computational Linguistics* 7. Publisher: MIT Press, Cambridge Massachussets, pp. 625–641. DOI: 10.1162/tacl_a_00290. URL: https://aclanthology.org/Q19-1040 (visited on 2023-03-09).

Warstadt, Alex, Amanpreet Singh, and Davide Fiocco (Dec. 20, 2018). *CoLA Baselines*. original-date: 2018-02-26T02:10:43Z. URL: https://github.com/nyu-mll/CoLA-baselines (visited on 2023-03-06).

Yi, David K., James V. Bruno, Jiayu Han, and Peter Zukerman (Sept. 11, 2022a). *analyzing_verb_alternation_plms*. original-date: 2022-04-03T19:56:16Z. URL: https://github.com/kvah/analyzing_verb_alternations_plms (visited on 2023-05-09).

Yi, David K., James V. Bruno, Jiayu Han, Peter Zukerman, and Shane Steinert-Threlkeld (Dec. 8, 2022b). "Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models". In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. BlackboxNLP 2022. Ed. by Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 142–152. DOI: 10.18653/v1/2022.blackboxnlp-1.12. URL: https://aclanthology.org/2022.blackboxnlp-1.12 (visited on 2024-06-30).

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen (Nov. 24, 2023). *A Survey of Large Language Models*. DOI:

10.48550/arXiv.2303.18223. arXiv: 2303.18223[cs]. URL: http://arxiv.org/abs/2303.18223 (visited on 2024-07-14).

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27. URL: https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html (visited on 2024-07-12).

A Glossary

The columns in LaVA (see Table 2.1) are based on Levin (1993). These frames are not standalone, but should be combined to paint the full picture. The following tables show how the data should be entered and read. The left of each table shows all possible label combinations found in LaVA, with the corresponding verb frames (see Table 1.1) on the right-hand side. 'Yes' indicates that the verb is listed in Levin (1993) as participating in that particular frame, while 'no' means that it is listed as non-participating. A blank space means that the verb was not mentioned with relation to that frame.

sl	sl_noloc	sl_nowith	*with*	Locative
0	0	0	no	no
0	1	0	yes	no
0	0	1	no	yes
1	0	0	yes	yes

TABLE A.1: SPRAY-LOAD alternation.

inch	non_inch	Inchoative	Causative
0	x	no	
1	0	yes	yes
0	1	no	yes
x	x		

TABLE A.2: CAUSATIVE-INCHOATIVE alternation.

there	non_there	There	No-There
0	1	no	yes
0	x	no	
1	0	yes	yes
x	x		

TABLE A.3: *there*-INSERTION alternation.

dat_both	dative_to	dative_do	Preposition	Double-Object
0	0	0	no	no
0	x	0		no
1	0	0	yes	yes
0	1	0	yes	no
0	0	1	no	yes

TABLE A.4: DATIVE alternation.

refl_op	refl_only	Non-Reflexive	Reflexive
0	0	no	no
0	1	no	yes
1	0	yes	
1	1	ambiguous (?)	

TABLE A.5: UNDERSTOOD-OBJECT alternation. The verbs ‘flex’ and ‘wave’ have a ‘1’ in both columns. This is likely an error.

B All results

	MCC				Accuracy				
	CoLA	GloVe	Baseline	Experiment	CoLA	GloVe	Baseline	Majority BL	Experiment
Causative-Inchoative									
Inchoative	0.555	0.672	0.741	0.732	0.810	0.855	0.885	0.680	0.884
Causative	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000
Dative									
Preposition	0.320	0.000	0.701	0.608	0.866	0.850	0.925	0.859	0.910
Double-Object	0.482	0.000	0.614	0.561	0.883	0.853	0.913	0.863	0.906
Spray-Load									
With	0.645	0.585	0.633	0.652	0.858	0.839	0.848	0.707	0.858
Locative	0.253	0.145	0.525	0.548	0.729	0.734	0.834	0.745	0.838
there-Insertion									
No-There	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000
There	0.459	0.536	0.593	0.548	0.843	0.858	0.876	0.762	0.847
Understood-Object									
Reflexive	0.000	0.000	0.466	0.390	0.977	0.976	0.867	0.853	0.866
Non-Reflexive	0.219	0.300	0.563	0.599	0.790	0.732	0.984	0.981	0.988
Resultative Construction									
Resultative	-	-	0.615	0.543	-	-	0.905	0.781	0.855
Selected	-	-	0.591	0.549	-	-	0.841	0.632	0.794
Non-Selected	-	-	0.560	0.670	-	-	0.783	0.545	0.836

Table B.1: Results from the experiment. The results from CoLA and Glove are from Kann et al. (2019), but as they are not obtained using the extended LaVA dataset, one can only draw superficial conclusions from those values. 'Majority BL' is short for 'majority baseline'.

Frame	Accuracy Δ	Majority BL Δ	MCC Δ
Inchoative	-0.000623	0.015942	-0.008664
Causative	0.000000	0.000000	0.000000
Preposition	-0.013230	0.006434	-0.086188
Double-Object	-0.004941	0.006892	-0.047051
With	0.003777	0.001707	0.005082
Locative	-0.007732	-0.004344	-0.014711
No-There	0.000000	0.000000	0.000000
There	-0.029290	-0.030936	-0.044903
Reflexive	0.004056	0.020401	-0.054355
Non-Reflexive	0.001866	0.002565	-0.043961
Resultative	-0.049504	-0.063552	-0.071837
Selected	-0.046462	-0.085038	-0.042219
Non-Selected	0.053755	-0.005270	0.110372

Table B.2: Difference between the baseline and the experiment.

C Errors in LaVA

C.1 Verbs

In the original dataset, two verbs are misspelt: the past tense of the verb *spread* was erroneously conjugated as *spreaded*, and *inundated* is entered as *innundated*. While these are only two non-existent word out of 516/586, we wonder what the impact of these misspellings are on the results. Since BERT has likely never encountered these words during pre-training, their embeddings are probably not representative of the actual words *spread* and *inundated*. Since both verbs have similar labels (0's and 1's in the same alternations), we expect it will have a minor effect on the frames belonging to the DATIVE, SPRAY-LOAD, and UNDERSTOOD-OBJECT alternations.

C.2 BNC

Kann et al. (2019) use embeddings trained on the 100M token British National Corpus (BNC). This corpus is restricted to British English words only (Leech, 1993), yet the spelling used in LaVA is US English. Furthermore, they replaced all words outside the 100k most frequent words in the BNC with `<unk>`. After an investigation, we observed that 52 of the 516 verbs in LaVA are not in the aforementioned 100k verbs, and 11 of 516 are not even in the BNC *at all* (Kilgarriff, 1996). Six of those 11 are the 'verbs' consisting of two words (see section 3.3.3). The other words are *coldcreamed, innundated, interweaved, smoldered* and *spritzed*. The absence in the BNC would only affect the results for the CoLA embeddings.