Adverbs, Surprisingly

Dmitry Nikolaev[1] **Collin F. Baker**[2] **Miriam R.L. Petruck**[3] **Sebastian Padó**[1]
[1] IMS, University of Stuttgart, Germany
[2] International Computer Science Institute, Berkeley, USA
[3] FrameNet
Corresponding Author: dnikolaev@fastmail.com

Abstract

This paper begins with the premise that adverbs are neglected in computational linguistics. This view derives from two analyses: a literature review and a novel adverb dataset to probe a state-of-the-art language model, thereby uncovering systematic gaps in accounts for adverb meaning. We suggest that using Frame Semantics for characterizing word meaning, as in FrameNet, provides a promising approach to adverb analysis, given its ability to describe ambiguity, semantic roles, and null instantiation.

1 Introduction

Adverbs are the part of speech (POS) that has seen the least attention in (computational) linguistics, likely due to its challenging nature (Conlon and Evens, 1992). As Huddleston and Pullum (2002, 563) state, "the adverb is a [. . .] residual category [. . .] to which words are assigned if they do not satisfy the more specific criteria for nouns, verbs, adjectives, prepositions, and conjunctions."

Syntactically, they modify many POSs, except nouns (*eat porridge quickly, hardly noticeable*), or even complete clauses (*Probably, I'll come tomorrow*). They are semantically varied (Thomason and Stalnaker, 1973), ranging from intensifiers/modifiers (*absolutely*, *beautifully*) to temporal and spatial specifications (*yesterday, forward*), to so-called *speaker-oriented adverbs* yielding inferences about speaker attitudes, beliefs, and evaluations. Finally, adverbs can occupy different positions in sentences, creating complex issues of scoping and ambiguity (Alexiadou, 2004; Payne et al., 2010). Consider the following sentences:[1]

(1) a. Happily, they watched TV until dinner.
b. They happily watched TV until dinner.
c. They watched TV happily until dinner.
d. They watched TV until dinner happily.

[1] Huddleston and Pullum (2002, 575)

While language users tend to interpret Ex. 1b–1d as describing the TV watchers' mental state, Ex. 1a is ambiguous and can also be read as a positive evaluation of the situation by the speaker.

In sum, adverbs provide crucial information not just about the where and how of events, but also about attitudes and evaluations. However, relatively little research on adverbs exists in computational linguistics, although lexical factors are generally recognized as central for many NLP tasks (Berger et al., 2000). Lexical information is generally represented either in online dictionaries or by embeddings extracted from corpora (Turney and Pantel, 2010; Devlin et al., 2019; Peters et al., 2018). As a dictionary, WordNet (Miller et al., 1990) lists adverbs but only provides a relatively impoverished account, while lexicons for sentiment analysis (Benamara et al., 2007; Dragut and Fellbaum, 2014) and hedging detection (Jeon and Choe, 2009; Islam et al., 2020) only consider specific subtypes of adverbs as to how they modulate the intensity of adjectives. On the distributional side, adverbs have been considered from a derviational perspective (Lazaridou et al., 2013); yet, they are rarely scrutinized in detail. Among the standard benchmarks, only GLUE (Wang et al., 2018) and BLiMP (Warstadt et al., 2020) cover adverbs, and then only marginally. The same is true of approaches that combine dictionaries and embeddings (Faruqui et al., 2015). As a consequence, SOTA language models consistently struggle with adverb meaning, as Section 2.2 will demonstrate empirically.

This paper argues that Frame Semantics (Fillmore, 1985), as realized in FrameNet (FN) (Ruppenhofer et al., 2016), provides an efficacious framework to articulate the relevant aspects of adverb meaning. Specifically, as Ex. 1 illustrates, lexical ambiguity is captured in terms of frame ambiguity. Moreover, inferences about the arguments of adverbs, typically filled by the speaker and the lexical unit that the adverb modifies, can be cap-

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 512–526
July 13-14, 2023 ©2023 Association for Computational Linguistics

tured and characterized via the frame elements (i.e. semantic roles) of the frame. Notably, FrameNet mechanisms will account for null-instantiated roles, allowing it to hint at unexpressed content in cases like Example 2b (v. Section 4.2 for details).

(2) a. [SPEAKER The Minister] **reported** [MESSAGE that the cost had exploded].
b. [MESSAGE The cost had] **reportedly** [MESSAGE exploded].

In such cases specifically, FrameNet considerations of frame element realization help to explain the absence of the SPEAKER semantic role in 2b.

Plan of the Paper. Section 2 defines the scope of this paper (speaker-oriented adverbs) and shows the lack of accounts for adverbs in NLP through a literature review. Section 3 presents a probing dataset for speaker-oriented adverbs on the basis of which it demonstrates empirically that current large language models do not provide accounts for adverb meaning. Section 4 provides general background information on FrameNet, gives details on the framework's approach to the description of adverb meaning, and suggests its use to improve NLP models. Section 5 concludes the paper.

2 Scope and Motivation

2.1 Scope

Given the variety and heterogeneity of adverbs, we restrict the empirical scope of this paper to a subclass of them – even though we believe that the conceptual points apply to adverbs generally. We focus on *speaker-oriented adverbs* (Ernst, 2009). This broad class of adverbs, itself comprises several subtypes brought together by their giving rise to a range of inferences about attitudes and beliefs of the speaker, such as epistemic beliefs (Ex. 3), evaluations (Ex. 1 and 4), and speech acts (Ex. 5):

(3) Peter says: "Paul is **certainly** right". $\models$ Peter is certain that Paul is right.

(4) Peter says: "**Unfortunately**, Paul arrived". $\models$ Peter is unhappy that Paul arrived.

(5) Peter says: "**Frankly**, Paul annoys me." $\models$ Peter voices his frank opinion.

Structurally, these entailments are similar to entailments that arise from implicative verbs (Karttunen, 1971). As sources of information about how speakers assess states of affairs, they are highly relevant for tasks like opinion mining (Pang and Lee, 2008) and stance detection (Thomas et al., 2006). However, while implicative verbs have received considerable attention in the context of textual entailment (Karttunen, 2012; Lotan et al., 2013), speaker-oriented adverbs have not.

2.2 Treatment of Adverbs in Computational Linguistics

This section summarizes work on adverbs in computational linguistics in the four most relevant areas: WordNets, applications, distributional modeling, and semantic annotation. Section 3 covers large language models separately.

WordNets. Princeton WordNet (WN, version 1.3) (Miller et al., 1990) covers about 4,500 English adverbs, comprising both single words and adverbial multi-word expressions like *a priori*. The information recorded includes senses (although most adverbs are monosemous) and semantic relations: almost all single-word adverbs are linked to the adjectives from which they are derived,and some adverbs have antonyms. However, WN has no information on the adverbs' syntactic or semantic behavior. The approach of corresponding WordNet resources varies substantially: GermaNet, for German, does not treat adverbs at all (Hamp and Feldweg, 1997). In contrast, plWordNet (Maziarz et al., 2016) provides a considerably richer description of adverbs, notably regarding lexical relations, but is only available for Polish.

NLP applications. Apparently, sentiment and emotion analysis are the NLP applications that have paid the most attention to adverbs (Benamara et al., 2007; Dragut and Fellbaum, 2014; Chauhan et al., 2020). Hedge detection, that is, the recognition of expressions that modulate speaker confidence in their statements boasts additional work on adverbs (Jeon and Choe, 2009; Islam et al., 2020). However, these studies, are generally limited to two specific subtypes: scalar adverbs that modify sentiment strength (intensifiers/minimizers: *very/hardly nice*) and adverbs that modify confidence (*certainly/apparently*). Haider et al. (2021) also considers locative and temporal adverbs. Confidence-modifying adverbs form a subtype of the speaker-oriented adverbs addressed here, but existing studies do not offer a general account of these adverbs beyond the requirements of specific tasks.

Studies on structured sentiment and emotion analysis (Barnes et al., 2021; Kim and Klinger, 2018) assume a different perspective. These works concentrate on defining and modeling the relations between sentiment- and emotion- introducing expressions and their semantic arguments, such as the experiencer of the affect and its target. As the comparison with Example 2 shows, these relations are at times tied to adverb meanings. However, we are not aware of studies in this area that deal specifically with adverbs.

Distributional modeling. A number of studies investigated the interplay between word embeddings and morphology, analyzing similarity by parts of speech (Cotterell and Schütze, 2015) or investigating meaning shifts corresponding to morphological derivation (Lazaridou et al., 2013; Padó et al., 2016). Typically, these studies include adverbs, and not surprisingly find that adverbs behave highly inconsistently.

Semantic annotation. In principle, frameworks for the annotation of (semantic) argument structure are promising sources for information about adverb meaning, but they differ widely in the information that they offer. The PropBank (Palmer et al., 2005) annotation scheme offers a range of modifier roles (ARGM) for the annotation of modifiers, including adverbs. However, the most fitting of these roles, ARGM-ADV, is a "catch-all" category. In addition, the PropBank analysis does not treat adverbs as predicates in their own right and does not assign roles to them. Thus, ***fortunately****, she accepted* and ***even*** *she accepted* would receive the same analysis.

In contrast, UCCA (Abend and Rappoport, 2013) explicitly splits adverbs into adverbial modifiers proper (D) and ground elements (G), where the latter expresses the speaker's attitude toward the event. However, UCCA does not make the structural relations explicit either.

AMR (Banarescu et al., 2013) offers a more nuanced approach: many adverbs are mapped to their underlying predicates and endowed with complete argument structure,[2] while others are interpreted as degree, manner, or time modifiers. However, no provision exists in the representation for speaker-oriented adverbs. To illustrate, the AMR annotation of *thankfully, she accepted the present* either treats the adverb as describing a general state of affairs (*it is good that she accepted*) or simply omits it.

Finally, Frame Semantics (Fillmore, 1985) offers the conceptual infrastructure to improve on these treatments and avoid their limitations. Section 4 provides justification of this understanding.

3 Case Study: Modeling Adverb Meaning as Natural Language Inference

One possibility, so far not mentioned, is that the knowledge inherent in large neural language models might provide a sufficient account of the meaning of (speaker-oriented) adverbs. In that case, at least from the NLP perspective, no (new) specific treatment would be required. However, this state of affairs is not the case, as we show below.

3.1 Creating Probing Datasets

To operationalize "a sufficient account," we ask language models to distinguish between valid and invalid inferences along the lines of Examples 3–5. As input data, we constructed probing examples with inferences for speaker-oriented adverbs.

We examined four classes of adverbs, motivated by current FrameNet frames containing adverbs (see Section 4.3 for details). These are: likelihood adverbs (e.g. *undoubtedly*, *probably*); unattributed-information adverbs (*reportedly*, *allegedly*, *supposedly*); degree adverbs (*at least*, *approximately*); and obviousness adverbs (*blatantly*, *conspicuously*).

We built the datasets from combinations of premises and hypotheses containing such adverbs, formulated as templates with sets of fillers for the adverbs and different participant positions. In this manner, we assessed the LM's capabilities irrespective of specific word choice. We paired each premise with two to four unambiguous hypotheses depending on the adverb class. The premise either implies or contradicts the hypothesis. Table 1 shows an example. Hypothesis 1 negates the premise and constitutes a contradiction. Hypothesis 2 is a valid inference about speaker evaluation; and Hypothesis 3 is a valid inference about the uncertainty inherent in the premise.

We report studies on two datasets with different emphases. We designed the first to be *naturalistic*, based on existing sentences for adverbs in FrameNet. Given the limited size of this dataset, we also created a larger *synthetic* dataset with simpler, more varied, sentences. The Appendix lists full details on both datasets.

Naturalistic Dataset. As stated, we created this dataset based on sentences in the FrameNet database containing adverbs of the four classes enumerated above. We "templatized" the sentences

[2]For example, AMR treats *sing* in *sing beautifully* as the first argument of `beautiful-02`.

Premise	The celebration had been postponed, **ostensibly** because of the Gulf War
Hyp 1	The Gulf War **ostensibly** had no effect on the celebration (CONTRADICTION)
Hyp 2	Someone said that the celebration was postponed because of the Gulf War (ENTAILMENT)
Hyp 3	The Gulf War may have had no effect on the celebration (ENTAILMENT)

Table 1: Naturalistic dataset: Probing items

Premise	SUBJ1 said that SUBJ2 **ADV** opposed ACTION
Hyp 1	SUBJ1 said that SUBJ2 may have opposed ACTION (ENTAILMENT)
Hyp 2	SUBJ1 is not sure that SUBJ2 opposed ACTION (ENTAILMENT)
Hyp 3	SUBJ1 is sure that SUBJ2 opposed ACTION (CONTRADICTION)
Hyp 4	SUBJ1 is sure that SUBJ2 did not support ACTION (CONTRADICTION)

Table 2: Synthetic dataset: Probing items

by treating the position of the adverb as a slot that can be filled by all semantically congruent adverbs from the respective class. In sentences where the subject is a personal name, we also treated the subject position as a slot, which we filled with twenty female and male names popular in the United States. Because the low number of sentences of the each type in the FrameNet database, and most templates have only one slot, viz. the adverb, the size of this dataset is limited. See Table 3 for example counts by adverb class.

Synthetic Dataset. The goal of this dataset was to test if the performance of the model is robust with regard to the replacement of the main-event description and varying syntactic complexity of the premises and hypotheses. It covers three of the four adverb classes: unattributed-information, degree, and obviousness, where the templates from the first dataset were most restricted. In these templates, subjects are always exchangeable. In addition, we also varied the description of the main action or relation described the sentence.

Table 2 shows the template set for unattributed-information adverbs. The set of adverbs for this class comprises *reportedly*, *allegedly*, *supposedly*, *apparently*, and *ostensibly*. Fillers of the ACTION slot include both gerund phrases (e.g. *selling the house*) and noun phrases (e.g. *the wedding*). Entailments and contradictions are produced in pairs. For entailments, we test two valid inferences triggered by the adverb. For contradictions, we test embedded clauses with and without negation. Table 5 shows the example count for each input type.

3.2 Probing Setup: NLI models

Arguably the best match for these types of datasets are the family of language models optimized for the task of natural-language inference (Storks et al., 2019). Concretely, we evaluated the series of NLI models released by Nie et al. (2020), the SNLI or Stanford Natural Language Inference models. These models carry out a three-way classification between ENTAILMENT, CONTRADICTION, and NEUTRAL. The author fine-tuned their models on a data set created in an iterative, adversarial, human-in-the-loop fashion, designed to remedy the shortcomings of previous NLI datasets (Belinkov et al., 2019). Preliminary experiments with different available base architectures (RoBERTa, ALBERT, BART, ELECTRA, and XLNet) showed that RoBERTa-large[3] was the best-performing variant. Thus, we used this model for evaluations. We used our probing datasets solely for evaluation, not for further fine-tuning.

For analysis, we checked the labels that the model predicted with their corresponding probabilities. In several cases, we performed additional tests to verify whether the adverbs or other properties of the sentence determined the model predictions.

3.3 Evaluation on a Naturalistic Dataset

3.3.1 Overall results

Table 3 shows overall results of the SNLI model on the naturalistic dataset for the four adverb classes. The adverb classes are not strictly comparable because they are represented by different input sentences (as described above), which include all types of lexical and syntactic confounds. Nevertheless, our experiments showed two consistent results: (i) the model cannot correctly draw inferences based on some set of adverbs on which it fails; (ii) the presence of adverbs increases the difficulty for the model to draw correct inferences in general. What follows is a survey of the evidence for these two claims.

3.3.2 Failure to Understand Adverbs

Degree adverbs. The model does not understand that *at least as big* is incompatible with *smaller*.

[3] ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

Adverb class	Error rate (%)	# sentences
Likelihood	2	5,880
Unattributed information	6	90
Degree	25	35
Obviousness	23	16

Table 3: Naturalistic dataset: SNLI model error rates by adverb class

While it correctly labels the pair *Lantau covers nearly twice the area of Hong Kong Island – Lantau is at least as big as Hong Kong Island* as ENTAILMENT and the same premise with *Lantau is much smaller than Hong Kong Island* as CONTRADICTION, it considers that this premise also entails *Hong Kong Island is at least as big as Lantau*, which is also a straightforward contradiction.

The quantifier–adverb combination *almost every* constitutes another weak point of the model. While it correctly labels the pair ***Almost all** assignments are challenging in different ways* vs. *Most of the assignments are difficult*, it labels ***Almost every** assignment is a challenge in a different way* vs. the same as NEUTRAL.[4]

Unattributed-information adverbs. The correct analysis of these adverbs is subtle since valid inferences may be expressed in ways that differ from the premise both lexically and syntactically.

Sometimes the model answers incorrectly with extremely high confidence. The example from Table 1 is a case in point. *The Gulf War ostensibly had no effect on the celebration* is always correctly labeled as CONTRADICTION. The *Someone said...* hypothesis is also correctly labelled as ENTAILMENT with **any** adverb in the premise. Strikingly, the model gives the same result when the adverb is omitted. This suggests that the model does not take the adverb in the premise into account.

The experiments with Hypothesis 3 (cf. Table 1) corroborated that understanding: regardless of the combination of the adverb in the premise and the hypothesis, the model confidently marks the pair as CONTRADICTION or NEUTRAL with almost zero probability attached to the prediction of ENTAILMENT. This finding shows that while the model may be able to draw a positive inference from the hearsay adverb (the reported event may have happened), it is completely unaware of the possibility of the negative inference, i.e. that the reported event may not have taken place: 12 times out of 16, the model confidently predicts CONTRADICTION.

[4]The model answers correctly only when there is a larger lexical overlap, as in *Most of the assignments are challenging.*

3.3.3 Adverbs Complicate Inference

In another analysis, we investigate the impact of the sentences' structural complexity on prediction quality. We frequently found that the model correctly inferred when the hypothesis is structurally simple or no adverb is given, but failed when the hypothesis had an embedded clause and the premise had an adverb. Table 4 shows a concrete example, which permits three observations:

1. The model is sensitive to whether the hypothesis contains an embedded clause: the confidence for the correct prediction drops from $\approx$1 to $\approx$0.8 for all verbs in the no-adverb case.
2. The presence of the adverb is not noticeable with structurally simple hypotheses: the confidence in the correct answer remains >0.9.
3. The combination of an adverb and an embedded clause can derail the model – paradoxically most so for the verb *support*, where the model was most confident without an adverb.

Furthermore, note that an adverb can force the model to change its decision even in the presence of a strong lexical cue. Given the hypothesis *The students were obviously drunk*, the model correctly identifies *The students abhor/forswore/renounced alcohol* as CONTRADICTION. While the model labels *The students abjured alcohol* as ENTAILMENT, (perhaps) because of an incorrect analysis of the verb, when we change the hypothesis to *The students were **conspicuously** drunk*, the model confidently and correctly labels *The students abjured alcohol* as CONTRADICTION.

3.4 Evaluation on a Synthetic Dataset

The results for the application of same model on the larger synthetic dataset are shown in Table 5. They demonstrate that in general the task of drawing correct inferences from adverbs is very difficult for the model. Instead, the model tends to consistently predict the same relation (entailment / neutral / contradiction) for all sentences for an adverb class. It is able to correctly predict inference for the quantity degree class (*at least two dozen people* $\models$ *many people* and $\not\models$ *nobody*). However, even syntactically trivial entailments and contradictions in other classes lead to systematic failures. E.g., while the model can correctly identify the inference *James said that Mary reportedly opposed the wedding* $\models$ *James said that Mary may have opposed the wed-*

Verb	Prediction	Hypothesis	obviously	clearly	publicly	blatantly	no adverb
aid	Entailment	Simple	0.94	0.94	0.95	0.96	0.97
		Complex	0.60	0.62	0.70	0.71	0.85
	Neutral	Simple	0.05	0.05	0.05	0.04	0.02
		Complex	0.39	0.38	0.29	0.27	0.15
help	Entailment	Simple	0.92	0.92	0.92	0.95	0.97
		Complex	0.53	0.52	0.58	0.61	0.77
	Neutral	Simple	0.07	0.08	0.08	0.05	0.03
		Complex	0.47	0.47	0.41	0.38	0.22
support	Entailment	Simple	0.99	0.99	0.99	0.99	0.99
		Complex	0.41	0.43	0.57	0.39	0.85
	Neutral	Simple	0.01	0.01	0.01	0.01	0
		Complex	**0.55**	**0.53**	0.40	**0.40**	0.15

Table 4: Prediction of NLI model given *Castro ADV backed the rebels* as premise and *Castro VERBed the rebels* or *Castro tried to VERB the rebels* as hypothesis (*simple* and *complex* respectively). Boldface indicates wrong model predictions; underline indicates "borderline correct" cases where an incorrect label received a probability $> 40\%$.

Semantic type	Test	Entailment	Neutral	Contradiction	Error rate (%)	# sentences
Unattributed information	Entailment 1	70,188	12	0	≈ 0	70,200
	Entailment 2	134	70,066	0	≈ 100	70,200
	Contradiction 1	7,940	62,260	0	100	70,200
	Contradiction 2	567	69,633	0	100	70,200
Degree (properties of people)	Entailment	31,200	0	0	0	31,200
	Contradiction	12,390	3,980	14,830	52	31,200
Degree (properties of objects)	Entailment	840	0	0	0	840
	Contradiction	547	0	293	65	840
Degree (quantities)	Entailment	38,400	0	0	0	38,400
	Contradiction	0	0	38,400	0	38,400
Obviousness	Entailment 1	54,600	0	0	0	54,600
	Entailment 2	33,217	21,383	0	39	54,600
	Contradiction 1	61	0	54,539	≈ 0	54,600
	Contradiction 2	0	1,615	52,985	3	54,600

Table 5: Synthetic dataset: Model predictions (cells with correct predictions have gray background) for each template class and error rates.

ding, it fails on the entailment of the type *James* ***is not sure*** *that Mary opposed the wedding*.

Similarly, with obviousness adverbs, while the examples of the type *James blatantly criticized Mary* $\models$ *James disparaged Mary* are easy for the model, entailments like *James tried to disparage Mary* leads to near-chance performance. In the domain of adverb-modulated relations, while the model seems to do well on entailments (*James is at least twice as rich as Mary* $\models$ *James's net worth is at least as big as Mary's*), in fact it does not understand that the relation is not symmetric and therefore cannot correctly identify contradictions (*Mary's net worth is at least as big as James's*).

3.5 Discussion

Taken together, these experiments demonstrate systematic shortcomings in the ability of current large language models to account for adverb meaning, either glossing over them completely or making rather random inferences about their meaning. Arguably, this study only looked at a specific type of language model and other types of language models would fare better. However, converging evidence from the literature exists.

For instance, Nikolaev and Padó (2023) analyzed sentence transformers, which might be expected to provide the most nuanced understanding of adverbs. Instead, the study found that the sentences' main participants (subjects and objects) primarily determine the semantic similarity of sentence pairs, which is largely independent of adverbs. The paper argues that this behavior arises from the structure of the training data for sentence transformers (online conversations, duplicate questions on WikiAnswers), where sentence pairs labelled as semantically similar often have similar sets of main participants (subjects and objects) and can vary widely in other respects.

If a similar bias is at play in the NLI models in the present study, creating larger, richer training sets that involve adverbs in a systematic manner is a way forward. However, given the relative scarcity of adverbs and their complex behavior (cf. Section 1), it seems unlikely that this effect will emerge naturally by pre-training on ever larger datasets. Instead, the evidence provided here indicates that adverb data must be created intentionally. The following section outlines a proposal to do so.

4 Describing Adverbs in FrameNet

This section will provide a brief background to FrameNet (Section 4.1), show how FrameNet can be useful for the analysis of adverbs (Section 4.2), survey the data on adverbs contained in the current version of the dataset (Section 4.3), and propose concrete directions for next steps (Section 4.4).

4.1 Background to FrameNet

FrameNet (*FN*, Ruppenhofer et al. 2016) is a research and resource-development project in corpus-based computational lexicography grounded in the theory of *Frame Semantics* (Fillmore, 1985).

At the heart of the work is the *semantic frame*, a script-like knowledge structure that facilitates inferencing within and across events, situations, states-of-affairs, relations, and objects. FN defines a semantic frame in terms of its *frame elements* (*FEs*), or participants (and other concepts) in the scene that the frame captures; a *lexical unit* (LU) is a pairing of a lemma and a frame, characterizing that LU in terms of the frame that it evokes. FN frames may include more than one POS, and FrameNet does not claim that the LUs of a frame are synonymous, merely that they are semantically similar in referring to the same situation. Additionally, FN distinguishes between core FEs and non-core FEs; the former uniquely define a frame and the later identify concepts that characterize events or situations more generally, such as time and place. To illustrate, Example 6 shows annotation for the verb *BUY*, defined in the `Commerce_buy` frame, with the FEs BUYER, SELLER, GOODS, and MONEY.[5]

(6) [Chuck $_{\text{BUYER}}$] **BOUGHT** [a car $_{\text{GOODS}}$] [from Jerry $_{\text{SELLER}}$] [for $2,000 $_{\text{MONEY}}$]

FrameNet annotators label approximately 20 sentences for each LU in each frame; and automatic processes tabulate the results to produce *valence* descriptions, or *semantic-syntactic combinatorial possibilities* of each LU. These also include *null-instantiated* core FEs, i.e. FEs that uniquely define a frame, even when not realized linguistically. Such valence descriptions provide information about meaning-form mappings that are important for natural-language understanding. FrameNet data, or semantic parsers built from them, have proven useful for tasks such as recognizing paraphrases (Ellsworth and Janin, 2007), drawing inferences (Ben Aharon et al., 2010), machine translation (Zhai et al., 2013), question answering (Khashabi et al., 2018), or paraphrasing (Wang et al., 2019).

At present, the FrameNet database (Release 1.7) holds 1,224 frames, defined in terms of 10,478 frame-specific FEs, and 13,686 LUs. Of those lexical units, 61% have *lexicographic* annotation, i.e. annotation for one target lemma per sentence.

4.2 FrameNet for the Analysis of Adverbs

We now outline how the descriptive devices of FrameNet, as outlined in Section 4.1, can capture the relevant facts about adverb meaning and address the core challenges of adverb classes, ambiguity, inferences, and null instantiation of roles.

Frames. Since frame definitions encompass much of the meaning of each LU, many FN frames already offer fine-grained, semantically motivated descriptions of adverb classes. For example, the `Emotion_directed` frame captures the semantic similarity of *happy*, *happily*, *happiness*, *sad*, and *sadly* and offers a starting point for the description of emotion-related adverbs, by exploiting the fact that these adverbs evoke the same background knowledge as the corresponding LUs of other parts of speech (Ruppenhofer et al., 2016).

When a lemma is ambiguous, each sense gets mapped to a different frame; each mapping is a separate lexical unit (LU). For instance, Example 1 in Section 1 includes the lemma *happily*, which is ambiguous: In Example 1a, *happily* is defined in the `Luck` frame (along with *fortunately* and *luckily*). The definition of this frame indicates that there is someone, the PROTAGONIST, for whom a particular state of affairs is surprisingly good or

[5]This paper uses the following typographical conventions: frame names appear in `typewriter font`; FE names are in SMALL CAPS; and lexical units are in **BOLD CAPS.**

bad. But this sentence does not express the PROTAGONIST; this is a case of null instantiation or NI (see below for details). The other three sentences, Examples 1b–1d, illustrate *happily* in the `Emotion_directed` frame. This involves an emotional response of someone, the EXPERIENCER, to a stimulus, the STIMULUS FE (here, watching TV), which evokes the emotional response, specifically happiness (recoverable from the definition of the LU *happily*). In these examples, the EXPERIENCER is explicit, so no inference is required (although coreference resolution will be required to resolve the referent of *they*). Example 7 shows the annotations of the sentences in the `Luck` frame (Ex. 7a) and in the `Emotion_directed` frame (Ex. 7b):

(7) a. HAPPILY, [they watched TV until dinner STATE_OF_AFFAIRS] PROTAGONIST: NI.
b. [They EXPERIENCER] HAPPILY [watched TV until dinner STIMULUS].

Frame Elements. In FrameNet, frame elements are associated with (classes of) inferences (Chang et al., 2002). Such inferences can capture important aspects of adverb meaning, as we have shown in Section 2. The valence patterns for the two senses of *happily* shown above lead to different inferences via the two sets of frame elements:

`Luck`: A STATE_OF_AFFAIRS is evaluated as good (or bad) [...] for a particular PROTAGONIST.

`Emotion_directed`: An EXPERIENCER [feels or experiences] a particular emotional response to a STIMULUS or about a TOPIC.

While such natural language descriptions were traditionally hard to formalize, the recent advances in "prompting" language models (Shin et al., 2020) have reestablished natural language descriptions as sufficient in many conditions (cf. also our template-based probing dataset in Section 3).

Null instantiation. FrameNet annotates information about the conceptually required "core" semantic roles of a frame even if absent from the text. FN distinguishes three types of null instantiation, one licensed by a construction and the others licensed lexically. FrameNet includes approximately 55,700 NI labels in its annotations; and roughly one-quarter of these omissions are licensed constructionally, with the remaining 75% licensed lexically (Petruck, 2019).

This capability of FrameNet is particularly important for adverbs, notably speaker-oriented adverbs. By definition, these adverbs welcome inferences about the speaker, who is typically not realized unless the statement is part of reported speech or thought: *The father thought: "Happily they are all watching TV."*

Returning to Example 2 (above), 2a illustrates an instantiated SPEAKER and 2b illustrates a *null-instantiated* SPEAKER, a fact that FN records in its database. No other lexical resource used extensively in computational linguistics records such information.

4.3 Current Status of Adverbs in FrameNet

Currently, FrameNet (Release 1.7) contains 217 adverb LUs. Of these adverbs, 158 have annotation, with a total of 2,475 annotations of adverbs on sentences in the database, yielding a mean of 16 annotations per LU. However, like many linguistic phenomena, the annotations exhibit a highly skewed (Zipfian) distribution: 41 of the 158 LUs have only one annotation while nine have more than 50 annotations each. In line with its general principles, FrameNet chose not to define one single frame to capture all speaker-oriented adverbs, instead defining each such adverb according to the specific frame it evokes. At the same time, the class of speaker-oriented adverbs is arguably recoverable from the union of a set of frames all of which support inferences about the speaker by way of describing the speaker through a certain frame element. In this way, the existing frames and their annotations provide a suitable basis for creating data for this (and future) research.

Table 6 shows the four FrameNet frames used to suggest adverbs for the experiment described in Section 3 together with the adverbs listed, illustrative example sentences, and their definitions.

4.4 Next Steps

As the numbers show (Section 4.3), FrameNet has not attended to adverbs either. Perhaps this fact represents a principal incompatibility: the description of adverbs may not welcome using concepts that FN developed for traditional predicates with clear-cut valence. Yet, we believe that including adverbs in FrameNet both follows the spirit of what Fillmore (1985) called "semantics of understanding" and is in line with FrameNet practice. Still, it will require work on two principal levels: theoretical development and practical lexicographic analysis.

Frame name	Adverbial lexical units & example sentence	Definition
`Unattributed information`	*allegedly.adv, ostensibly.adv, purportedly.adv, reportedly.adv, supposedly.adv* **Ex.** One person was REPORTEDLY killed. . .	A speaker presents a REPORTED FACT as deriving from statements (made directly to them or to others) of third parties.
`Likelihood`	*certainly, likely, probably, possibly* **Ex.** This will LIKELY not be enough to stop. . .	This frame concerns the likelihood of a HYPOTHETICAL EVENT occurring, the only core frame element in the frame.
`Obviousness`	*audibly.adv, clearly.adv, evidently.adv, noticeably.adv, obviously.adv, visibly.adv* **Ex.** It is CLEARLY desirable to permit the gifted youngsters to flourish.	A PHENOMENON is portrayed in terms of the DEGREE of likelihood that it will be perceived and known, given the (usually implicit) EVIDENCE, PERCEIVER, and CIRCUMSTANCES in which it is considered.
`Degree`	*a little (bit).adv, a lot.adv, absolutely.adv, as hell.adv, far.adv, fully.adv, in part.adv, kind of.adv, so.adv, somewhat.adv, that.adv, totally.adv, very.adv, way.adv* **Ex.** I had ABSOLUTELY nothing to say.	LUs in this frame modify a GRADABLE ATTRIBUTE and describe intensities at the extreme positions on a scale.

Table 6: FrameNet Frames characterizing Speaker-Oriented Adverbs

At the theoretical level, the FrameNet approach has seen constant development over the 25 years of the project's existence. In initial verb-centered frames, nominals tended to fill FEs, with additional attributes realized as adverbs. Next, FN added deverbal nouns to frames, which largely take the same frame elements. To expand to other types of nouns, like natural kinds and artifacts, FrameNet broadened the concept of FE to encompass *qualia* such as substance or purpose (Pustejovsky, 1991). Layering the annotation of nouns as FEs of verbs, and modifiers of nouns as *their* FEs provided a richer semantic representation. Next, FrameNet included adjectives as frame-evoking elements, permitting generalizations over domains like speed or temperature. While most aspects of adverbs description are already present in FrameNet (cf. above), theoretical analysis must make precise the implications of annotating null instantiated adverbial frame elements at scale.

At the practical level, the time is ripe to add many more adverbs to appropriate existing frames and to create new frames for adverbs as needed. The principles of annotating naturally occurring text and extracting valence descriptions for LUs established on the other parts of speech carry over to adverbs. The combination of valence descriptions and annotated instances constitute essential inputs to characterize inferences.

Clearly, the more annotation, the better, but large-scale expert annotation is slow and resource-intensive. Using crowdsourcing, which permits parallelizing (thus, speeding up) annotation, is a possible mitigation. Fossati et al. (2013) and Feizabadi and Padó (2014) demonstrated success with crowdsourcing for frame-semantic annotation when the task is narrowed down appropriately. Substantial promise exists to extract adverb annotation automatically from comparable corpora (Roth and Frank, 2015) and paraphrasing models (Wang et al., 2019). Even for the core task of FrameNet analysis, defining frames, Ustalov et al. (2018) proposed automatic methods. Still, full automation remains hard, given concerns of quality and consistency.

5 Conclusion

Conlon and Evens (1992) stated that adverbs are under-researched in computational linguistics; this statement is still true. Adverbs have received attention only in two applications: sentiment analysis and hedging detection. The large language models used here show systematic gaps in capturing adverb meaning. The problem is **not** solved.

We propose that Frame Semantics, as embodied in FrameNet, along with improved techniques to mitigate the annotation effort to extend FN with new frames and annotations, can capture the meaning and implicatures of adverbs. Considering frames as lexical constructions (Fillmore, 2008), this proposal fits well with recent work to combine language models and construction grammar (Weissweiler et al., 2023).

Multiple ways exist for computational modeling to use such a resource, e.g., by extending the coverage of semantic role labellers to a larger range of adverbs, or by fine-tuning language models on large annotated datasets for which our probing dataset can serve as a blueprint.

Limitations

We only used English data in the study, so we cannot guarantee that the findings will generalize to other languages (cf. Bender 2019). The English NLI datasets are, as usual, larger than for other languages, so we should expect models targeting other languages to have worse performance. We do, however, believe that the challenges of adverbs are comparable in other languages, particularly in typologically similar languages.

Ethics Statement

The paper argues for a new approach to the treatment of adverbs in the development of resources and applications in NLP. We consider better understanding of language by computational models as not posing a significant societal risk in itself. The dataset used for the computational experiment in Section 3 was created based on the data contained in the publicly available FrameNet corpus and, as far as we are aware, does not contain sensitive elements. Implementation of our proposed methodology has the same risks as any data-driven approach in computational linguistics, but we assume that we cannot safeguard against its possible misuse due to its very general nature.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Artemis Alexiadou. 2004. Adverbs across frameworks. *Lingua*, 114(6):677–682.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Roni Ben Aharon, Idan Szpektor, and Ido Dagan. 2010. Generating entailment rules from FrameNet. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 241–246, Uppsala, Sweden. Association for Computational Linguistics.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and V. S. Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of ICWSM*.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 192–199, New York, NY, USA. Association for Computing Machinery.

Nancy Chang, Srini Narayanan, and Miriam R.L. Petruck. 2002. Putting frames in perspective. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

U.A. Chauhan, M.T. Afzal, A. Shahid, M. Abdar, M.E. Basiri, and X. Zhou. 2020. A comprehensive analysis of adverb types for mining user sentiments on Amazon product reviews. *World Wide Web*, 23:1811–1829.

Sumali Pin-Ngern Conlon and Martha Evens. 1992. Can computers handle adverbs? In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eduard Dragut and Christiane Fellbaum. 2014. The role of adverbs in sentiment analysis. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 38–41, Baltimore, MD, USA. Association for Computational Linguistics.

Michael Ellsworth and Adam Janin. 2007. Mutaphrase: Paraphrasing with FrameNet. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 143–150, Prague. Association for Computational Linguistics.

Thomas Ernst. 2009. Speaker-oriented adverbs. *Natural Language and Linguistic Theory*, 27:497–544.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Parvin Sadat Feizabadi and Sebastian Padó. 2014. Crowdsourcing annotation of non-local semantic roles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 226–230, Gothenburg, Sweden. Association for Computational Linguistics.

Charles J Fillmore. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica*, IV(2):222–254.

Charles J Fillmore. 2008. Border conflicts: Framenet meets Construction Grammar. In *Proceedings of the XIII. EURALEX international congress*, volume 4968, pages 49–68, Barcelona, Spain.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 742–747, Sofia, Bulgaria. Association for Computational Linguistics.

Sajjad Haider, Muhammad Tanvir Afzal, Muhammad Asif, Hermann Maurer, Awais Ahmad, and Abdelrahman Abuarqoub. 2021. Impact analysis of adverbs for sentiment classification on Twitter product reviews. *Concurrency and Computation: Practice and Experience*, 33(4):e4956. E4956 CPE-18-0194.R2.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Jumayel Islam, Lu Xiao, and Robert E. Mercer. 2020. A lexicon-based approach for detecting hedges in informal text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3109–3113, Marseille, France. European Language Resources Association.

Jieun Jeon and Jae-Woong Choe. 2009. A key word analysis of English intensifying adverbs in male and female speech in ICE-GB. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 210–219, Hong Kong. City University of Hong Kong.

Lauri Karttunen. 1971. Implicative verbs. *Language*, 47(2):340–358.

Lauri Karttunen. 2012. Simple and phrasal implicatives. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 124–131, Montréal, Canada. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526, Sofia, Bulgaria. Association for Computational Linguistics.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. TruthTeller: Annotating predicate truth. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–757, Atlanta, Georgia. Association for Computational Linguistics.

Marek Maziarz, Stan Szpakowicz, and Michal Kalinski. 2016. Adverbs in plWordNet: Theory and implementation. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 209–217, Bucharest, Romania. Global Wordnet Association.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on WordNet. *International Journal of Lexicography*, 3(4):235–312.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023. Representation biases in sentence transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3701–3716, Dubrovnik, Croatia. Association for Computational Linguistics.

Sebastian Padó, Aurélie Herbelot, Max Kisselew, and Jan Šnajder. 2016. Predictability of distributional semantics in derivational word formation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1285–1296, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

John Payne, Rodney Huddleston, and Geoffrey K. Pullum. 2010. The distribution and category status of adjectives and adverbs. *Word Structure*, 3(1):31–81.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Miriam R L Petruck. 2019. Meaning representation of null instantiated semantic roles in FrameNet. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 121–127, Florence, Italy. Association for Computational Linguistics.

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

Michael Roth and Anette Frank. 2015. Inducing implicit arguments from comparable texts: A framework and its applications. *Computational Linguistics*, 41(4):625–664.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. ICSI: Berkeley.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Richmond H. Thomason and Robert C. Stalnaker. 1973. A semantic theory of adverbs. *Linguistic Inquiry*, 4(2):195–220.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Dmitry Ustalov, Alexander Panchenko, Andrey Kutuzov, Chris Biemann, and Simone Paolo Ponzetto. 2018. Unsupervised semantic frame induction using triclustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 55–62, Melbourne, Australia. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. 2019. A task in a suit and a tie: paraphrase generation with semantic augmentation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Georgetown University Round Table Workshop on CxGs+NLP*.

Feifei Zhai, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2013. Handling ambiguities of bilingual predicate-argument structures for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1136, Sofia, Bulgaria. Association for Computational Linguistics.

A Details on the Naturalistic Dataset

The probing dataset includes a series of template classes. Each template class corresponds to an adverb class and contains several NLI templates with slots for adverbs and, when the structure permits it, also for the subject. In testing, we used all pairs of adverbs from the relevant class to instantiate the premise and the hypothesis. When a variable for subject exists in the premise, we used the same subject in the hypotheses.

A.1 Likelihood Adverbs

Adverbs: *undoubtedly*, *surely*, *positively*, *likely*, *certainly*, *definitely*, *totally*.

Fillers for the subject slot: *Barbara*, *Charles*, *David*, *Elizabeth*, *James*, *Jennifer*, *Jessica*, *John*, *Joseph*, *Karen*, *Linda*, *Mary*, *Michael*, *Patricia*, *Richard*, *Robert*, *Sarah*, *Susan*, *Thomas*, *William*.

1. **Premise:** *SUBJ is ADV gonna have to check it tomorrow afternoon again.*
   Entailment: *SUBJ is ADV going to have to check it again.*
   Contradiction: *SUBJ ADV won't need to check it again.*

2. **Premise:** *SUBJ can ADV find bargains in Tunis.*
   Entailment: *SUBJ will ADV find good deals in Tunis.*
   Contradiction: *SUBJ will ADV discover that everything is expensive in Tunis.*

3. **Premise:** *His friend, SUBJ, is ADV a foreigner.*
   Entailment: *SUBJ ADV is from another country.*
   Contradiction: *SUBJ ADV is a native here.*

A.2 Unattributed-information adverbs

Adverbs: *reportedly*, *allegedly*, *supposedly*, *apparently*, *ostensibly*.

1. **Premise:** *The German government ADV opposed the quotas.*
   Entailments: *The German government ADV was against the quotas*; *The German government may have supported the quotas.*
   Contradiction: *The German ADV supported more quotas.*

2. **Premise:** *The celebration had been postponed, ADV because of the Gulf War.*
   Entailments: *Someone said that the celebration was postponed because of the Gulf War*; *The Gulf War may have had no effect on the celebration.*
   Contradiction: *The Gulf War ADV had no effect on the celebration.*

A.3 Degree Adverbs

Adverbs: *at least*, *at a minimum*, *nearly*, *approximately*.

1. **Premise:** *Lantau covers ADV twice the area of Hong Kong Island.*
   Entailment: *Lantau is at least as big as Hong Kong Island.*
   Contradiction: *Hong Kong Island is at least as big as Lantau.*

2. **Premise:** *At the moment ADV 140 persons are working to curtail the fire.*
   Entailment: *Many people are fighting the fire.*
   Contradiction: *Nobody is fighting the fire.*

A.4 Obviousness Adverbs

Adverbs: *blatantly*, *obviously*, *clearly*, *ostentatiously*, *noticeably*, *visibly*, *conspicuously*.

1. **Premise:** *Castro ADV backed the rebels.*
   Entailments: *Castro helped the rebels*; *Castro tried to help the rebels.*
   Contradiction: *Castro tried to stop the rebels.*

2. **Premise:** *The students were ADV drunk.*
   Entailment: *The students were surely drinking too much.*
   Contradiction: *The students renounced alcohol.*

B Details on the Synthetic Dataset

B.1 Fillers for the human-subject slot

James, *Mary*, *Robert*, *Patricia*, *John*, *Jennifer*, *Michael*, *Linda*, *David*, *Elizabeth*, *William*, *Barbara*, *Richard*, *Susan*, *Joseph*, *Jessica*, *Thomas*, *Sarah*, *Charles*, *Karen*, *Li*, *Wei*, *Fang*, *Xiuying*, *Na*, *Priya*, *Rahul*, *Divya*, *Abhishek*, *Ishita*, *Melokuhle*, *Omphile*, *Iminathi*, *Lisakhanya*, *Lethabo*, *Ivaana*, *Malik*, *Pipaluk*, *Aputsiaq*, *Nivi*.

B.2 Unattributed-information adverbs

Adverbs: *reportedly*, *allegedly*, *supposedly*, *apparently*, *ostensibly*.

Actions: *the wedding*, *the marriage*, *buying the house*, *selling the car*, *moving away*, *staying in Canberra*, *delaying the funeral*, *the arrangement*, *the lawsuit*.

Premise: *SUBJ1 said that SUBJ2 ADV opposed ACTION.*

Entailments:

1. *SUBJ1 said that SUBJ2 may have opposed ACTION.*
2. *SUBJ1 is not sure that SUBJ2 opposed ACTION.*

Contradictions:

1. *SUBJ1 is sure that SUBJ2 opposed ACTION.*
2. *SUBJ1 is sure that SUBJ2 did not support ACTION.*

B.3 Degree adverbs

Adverbs: *at least*, *at a minimum*, *nearly*, *approximately*.

B.3.1 Properties of people

Properties: *net worth*, *knowledge*, *manners*, *fan base*, *culpability*.[6]

Adjectives:

- **Adjective 1**: *rich*, *erudite*, *polite*, *popular*, *guilty*.
- **Adjective 2**: *big*, *extensive*, *good*, *large*, *high*.

Premise: *SUBJ1 is ADV twice as ADJ1 as SUBJ2.*

Entailment: *SUBJ1's PROPERTY is/are at least as ADJ2 as SUBJ2's.*

Contradiction: *SUBJ2's PROPERTY is/are at least as ADJ2 as SUBJ1's.*

B.3.2 Properties of objects

Subjects: *the truck*, *the house*, *the hotel*, *the ship*, *the wagon*, *the car*, *the tree*.

Properties: *age*, *weight*, *height*, *width*, *price*.

Adjectives:

- **Adjective 1**: *old*, *heavy*, *tall*, *wide*, *expensive*.
- **Adjective 2**: *great*, *big*, *big*, *big*, *high*.

Premise: *SUBJ1 is ADV twice as ADJ1 as SUBJ2.*

Entailment: *The PROPERTY of SUBJ1 is at least as ADJ2 as that of SUBJ2.*

Contradiction: *The PROPERTY of SUBJ2 is at least as ADJ2 as that of SUBJ1.*

B.3.3 Quantities

Times: *at the moment*, *now*, *these days*, *this month*, *this week*.[7]

Numbers: *two dozen*, *thirty*, *fifty*, *140*.

Related-person groups: *friends*, *relatives*, *acquaintances*, *coworkers*.

Activities: *working on this*, *helping with the move*, *coming to visit us*.

Premise: *TIME ADV NUMBER of SUBJ's RELATED_PERSONS are ACTIVITY.*

Entailment: *Many people are ACTIVITY.*

Contradiction: *Nobody is ACTIVITY.*

B.4 Obviousness adverbs

Adverbs: *blatantly*, *obviously*, *clearly*, *ostentatiously*, *noticeably*, *visibly*, *conspicuously*.

Actions: [8]

- **Action 1**: *backed*, *supported*, *criticized*, *provoked*, *brainwashed*.

[6] Unlike in case with adverbs and subject-slot fillers, where all combinations are used, properties and adjectives in this and the next subclass are used in parallel. I.e., when the *i*'th adjective from the first list is used in the premise, the corresponding *i*'th property and adjective from the second list will be used in the hypotheses.

[7] Similarly to the two previous subclasses, times, numbers, activities, and related-person groups in this subclass are used in parallel. I.e., when the *i*'th time, number, related-person group, and activity are used in the premise, the corresponding *i*'th activity will be used in the hypotheses.

[8] Similarly to adjectives and properties in the case of degree adverbs above, actions of different types are used in parallel. I.e., when the *i*'th element from the first list is used in the premise, corresponding *i*'th elements from other lists will be used in the hypotheses.

- **Action 2, past**: *helped*, *encouraged*, *disparaged*, *incited*, *indoctrinated*.
- **Action 2, infinitive**: *help*, *encourage*, *disparage*, *incite*, *indoctrinate*.
- **Action 3, past**: *stopped*, *deterred*, *praised*, *calmed*, *deprogrammed*.
- **Action 3, infinitive**: *stop*, *deter*, *praise*, *calm*, *deprogram*.

Premise: *SUBJ1 ADV ACTION1 SUBJ2.*

Entailments:

1. *SUBJ1 ACTION2_PAST SUBJ2.*
2. *SUBJ1 tried to ACTION2_INF SUBJ2.*

Contradictions:

1. *SUBJ1 ACTION3_PAST SUBJ2.*
2. *SUBJ1 tried to ACTION3_INF SUBJ2.*