Quirk or Palmer: A Comparative Study of Modal Verb Frameworks with Annotated Datasets

Risako Owan, Maria Gini, and **Dongyeop Kang**
Computer Science and Engineering
University of Minnesota
{owan0002,gini,dongyeop}@umn.edu

Abstract

Modal verbs, such as *can*, *may*, and *must*, are commonly used in daily communication to convey the speaker's perspective related to the likelihood and/or mode of the proposition. They can differ greatly in meaning depending on how they're used and the context of a sentence (e.g. "They *must* work together." vs. "They *must* have worked together."). Despite their practical importance in natural language understanding, linguists have yet to agree on a single, prominent framework for the categorization of modal verb senses. This lack of agreement stems from high degrees of flexibility and polysemy from the modal verbs, making it more difficult for researchers to incorporate insights from this family of words into their work. As a tool to help navigate this issue, this work presents `MoVerb`, a dataset consisting of 27,240 annotations of modal verb senses over 4,540 utterances containing one or more sentences from social conversations. Each utterance is annotated by three annotators using two different theoretical frameworks (i.e., Quirk and Palmer) of modal verb senses. We observe that both frameworks have similar inter-annotator agreements, despite having a different number of sense labels (eight for Quirk and three for Palmer). With RoBERTa-based classifiers fine-tuned on `MoVerb`, we achieve F1 scores of 82.2 and 78.3 on Quirk and Palmer, respectively, showing that modal verb sense disambiguation is not a trivial task. [1]

1 Introduction

Modal verbs (also referred to as modal operators, modals, or modal auxiliaries (Imre, 2017)) convey important semantic information about a situation being described or the speaker's perspective related to the likelihood and/or mode of the proposition (Lyons, 1977; Quirk et al., 1985). Because of the widespread use of modal verbs in our daily lives, an accurate modeling of modal verb senses from context is essential for semantic understanding. For example, as modal verbs are often used with verbs that express one's personal state or stance, such as *admit, imagine*, and *resist* (Biber et al., 2002), we can utilize them for better speaker intention identification or sentiment analysis.

In both linguistics and NLP, however, there is no unifying consensus on how to organize these words (Table 1). One reason for this indeterminacy is their lack of a straightforward definition (Nuyts et al., 2010). Modal verbs have nuanced meanings, and their interpretation is often subjective. For example, if a speaker says, "I *can* go to the event today", it can refer to their ability to go to the event, the possibility that they might go to the event, or the fact that they obtained permission to go to the event. As such, categorizing modal verbs requires more attention than many other linguistic features, making the task challenging even for humans.

Two commonly used frameworks come from Quirk et al. (1985) and Palmer (1990). To compare these frameworks, we present a new dataset, `MoVerb`, containing 4540 annotated conversational English utterances with their modal verb categories. We chose the conversational domain since spoken, casual text is more flexible and nuanced compared to language from other domains and therefore could reap the most benefits from better modal verb classifications. To the best of our knowledge, this study provides the first empirical comparison of two modal verb frameworks with annotated datasets, evaluating the practicality of these different theoretical frameworks. Our study shows a clear inclination towards one of the two frameworks and quantitatively shows how humans struggle with the task.

Our main contributions are as follows:

- We collect `MoVerb`, an annotated conversational domain dataset containing two types of labels for modal verbs in 4540 English utter-

[1] Our dataset will be publicly available with our final version at https://github.com/minnesotanlp/moverb

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 183–199
December 6–7, 2023.

Reference	Modality Categories							
Kratzer (1991)	Epistemic			Deontic			Circumstantial	
Palmer (1986)	Epistemic			Deontic			Dynamic	
Quirk et al. (1985)	Possibility	Ability	Permission	Necessity[3]	Obligation[4]	Inference[5]	Prediction	Volition
Baker et al. (2010)	Requirement	Permissive	Success	Effort	Intention	Ability	Want	Belief
Ruppenhofer and Rehbein (2012)	Epistemic	Deontic	Dynamic	Optative	Concessive	Conditional		
Matthewson and Truckenbrodt (2018)	Root (Teleological Deontic Bouletic)				Epistemic (Inferential Reportative)			
Nissim et al. (2013)[6]	Epistemic (committment evidential)		Deontic (manipulative volition)		Dynamic (axiological appreciative apprehensional)			
Portner (2009)	Epistemic	Priority (Deontic Bouletic Teleological)			Dynamic (Volitional Quantificational)			

Table 1: A non-exhaustive list of past work on modality and the frameworks they use. Note that some linguists support two-tiered categorical frameworks by defining general categories that are further divided into subcategories.

ances. The dataset is split into two distinct parts. The first part consists of utterances with a single final label determined by majority voting and the second consisting of utterances with complete disagreement.[2]

- We observe the difficulty of annotating modal verbs even when based on solid theoretical frameworks. We discuss findings that suggest other causes of annotator disagreement besides a difference in sentence interpretation.
- We find a clear performance gap between the fine-tuned classifiers trained on different frameworks of data in `MoVerb`: 82.2 F1 on Quirk and 78.3 F1 on Palmer. Additionally, the classifier fine-tuned on Palmer's categories struggles when applied to a different domain.

2 Related work

There are numerous linguistic studies about modal verbs and their categorization (Quirk et al., 1985; Palmer, 1990; Lyons, 1977; Mindt, 2000; Kratzer, 2012; Morante and Sporleder, 2012; Aarts et al., 2021). However, despite attempts to reconcile them (Duran et al., 2021), widespread variation makes it unclear which framework would work best for specific NLP tasks. A dataset using multiple modal verb frameworks would help researchers experiment, but that dataset is yet to be built. To the best of our knowledge, there is no English dataset dedicated to the comparison of modal verb labeling.

Framework consistency is not the only thing lacking in modality datasets. Sources of modality can vary as well. In a multilingual corpus focusing on modality as a whole, Nissim et al. manually tag words and phrases representing modality. Due to the lack of emphasis on modal verbs, this dataset contains only 32 instances over 7 modal verbs: *will, might, can, may, would, could,* and *should* (Nissim et al., 2013). We argue that a dataset focusing on modal verbs is also necessary because of the ample complexities of modal verbs on their own.

Even datasets that do focus on modal verbs are not guaranteed to study the same set of words (Ruppenhofer and Rehbein, 2012; Marasović et al., 2016). Modal verbs in different domains, namely conversational and academic, have quite dissimilar distributions (Biber et al., 2002). In our cross-domain analysis, we utilized a dataset for subjectivity analysis in opinions and speculations from the news domain (Ruppenhofer and Rehbein, 2012; Wiebe et al., 2005). Ruppenhofer and Rehbein do not include *would* and *will* in their annotations, making their dataset challenging for analyzing conversational English. *Would* and *will* are 1st and 3rd when we rank modal verbs by their frequencies in spoken English (Mindt, 2000; Biber et al., 2002).

We note that there is a slight difference in our annotation frameworks. Ruppenhofer and Rehbein create a schema of their own, building off of work by Baker et al. (2010) and Palmer (1986). We do

[2]We acknowledge that majority voting has limitations when used in dataset creation and discuss this further in Section 5

[3]Logical Necessity

[4]Obligation/Compulsion

[5]Tentative Inference

[6]Nissim et al.'s work includes more categories on different dimensions, but we only show those comparable to the others in this table

UTTERANCES WITH COMPLETE AGREEMENT	ANNOTATOR 1	ANNOTATOR 2	ANNOTATOR 3
Usually moving your body helps but it depends on her situation... i *would* get a 2nd opinion!	volition	volition	volition
I bought a lottery ticket and have a feeling I *will* win.	prediction	prediction	prediction
That is really sweet of them. *Must* have been a big party.	necessity	necessity	necessity
I get it.. but you know life really is too short.. i think you *should* try to reach out! Do it!:)	obligation	obligation	obligation

UTTERANCES WITH COMPLETE DISAGREEMENT	ANNOTATOR 1	ANNOTATOR 2	ANNOTATOR 3
That *must* have been terrible. Were you okay?	inference	necessity	possibility
I am going to a drink and paint party tomorrow. It *should* be pretty fun!	inference	necessity	prediction
I am stressed by my blood test results that I *will* have tomorrow.	ability	necessity	prediction
I work remotely, I wish that you *could* do something like that as well.	ability	permission	possibility

Table 2: Annotation examples from MoVerb for complete agreement and disagreement among the three annotators. Note that *necessity* here refers to logical necessity, not social or physical necessities.

not use Baker et al.'s labels since we are more interested in applying traditional linguistic theories. However, we are still able to compare results since Palmer's categories make up 97.57% of the annotations in Ruppenhofer and Rehbein's dataset.

3 Potential Applications with Modal Verbs

There is some debate as to whether we should focus on modality as a whole since it can be expressed in other ways not limited to modal verbs (Nissim et al., 2013; Pyatkin et al., 2021). However, we argue that modal verbs alone offer enough complexity. There is untapped potential in improving the categorization of modal verbs, which could greatly enhance the performance of various downstream natural language processing (NLP) tasks.

Difficulty with modal verb understanding can cause confusion in semantic similarity tasks. Using a RoBERTa Hugging Face model (Liu et al., 2019) pretrained on the Microsoft Research Paraphrase Corpus (MRPC) subset of the General Language Understanding Evaluation (GLUE) dataset[7], we saw that the model was not able to reliably identify the unlikely interpretations for given sentences. For example, given the sentence, "My parents said I *can* go", the model would flag all following three as semantically equivalent by a score of at least 0.73: "My parents said I have the ability to go.', "My parents said I might go.", and "My parents said I have permission to go".[8]

As another example, we generated paraphrases for the Empathetic Dialogues dataset (Rashkin et al., 2019) using the T5 Parrot paraphraser (Damodaran, 2021) in the Hugging Face library.[9] This revealed that 1951 out of 2490 (78.35%) paraphrases created for 865 sentences[10] kept their original modal verbs. This suggests that being able to correctly identify and paraphrase the sense of a modal verb can greatly increase variety in paraphrasing.

4 Theoretical Frameworks

We use two labeling frameworks in our dataset annotations that we refer to as Quirk's categories and Palmer's categories.

- Quirk's categories consist of eight labels: *possibility*, *ability*, *permission*, *logical necessity* (abbrev. *necessity*), *obligation/compulsion* (abbrev. *obligation*), *tentative inference* (abbrev. *inference*), *prediction*, and *volition*. While the labels are self-explanatory, further descriptions can be found in Figures 5 and 6 of Appendix A.1. (Quirk et al., 1985)
- Palmer's categories consist of three labels: *deontic*, *epistemic*, and *dynamic*. A deontic modal verb influences a thought, action,

[7] textattack/roberta-base-MRPC

[8] 0.978, 0.732, and 0.988 respectively

[9] prithivida/parrot_paraphraser_on_T5

[10] We removed utterances with multiple sentences since paraphrase models will sometimes drop a sentence in an attempt to create a "new" paraphrase.

	POSSIBILITY	PREDICTION	INFERENCE	NECESSITY	ABILITY	VOLITION	PERMISSION	OBLIGATION
DEONTIC	50	21	22	27	42	31	22	288
EPISTEMIC	454	307	120	317	110	12	1	10
DYNAMIC	197	172	13	11	758	194	22	22

Table 3: The frequency distribution between Quirk's and Palmer's categories in MoVerb. This table shows that there is no clear mapping between the two frameworks, although there are common combinations (epistemic possibility, dynamic ability, etc.) that reveal overlapping categories.

	WILL	WOULD	SHOULD	MAY	MIGHT	MUST	COULD	CAN	TOTAL
POSSIBILITY	50	61	7	128	324	0	119	96	785 (0.22%)
ABILITY	14	24	0	0	0	1	302	657	998 (0.28%)
PERMISSION	2	4	4	19	1	0	10	12	52 (0.01%)
NECESSITY	7	12	13	0	0	334	3	1	370 (0.1%)
OBLIGATION	5	6	307	1	0	18	0	4	341 (0.1%)
INFERENCE	6	42	45	2	11	73	1	1	181 (0.05%)
PREDICTION	351	183	19	0	5	4	4	3	569 (0.16%)
VOLITION	129	92	11	3	6	1	6	6	254 (0.07%)
TOTAL	564 (16%)	424 (12%)	406 (11%)	153 (4%)	347 (10%)	431 (12%)	445 (13%)	780 (22%)	**3550**
EPISTEMIC	283	269	78	99	232	479	118	161	1719 (42%)
DEONTIC	32	65	437	25	18	35	27	52	691 (16.9%)
DYNAMIC	336	258	29	37	108	6	315	592	1681 (41.1%)
TOTAL	651 (16%)	592 (14%)	544 (13%)	161 (4%)	358 (9%)	520 (13%)	460 (11%)	805 (20%)	**4091**

Table 4: The breakdown of agreed-upon categories for each modal verb in MoVerb. Instances labeled *Unknown* by the annotators are excluded.

or event by giving permission, expressing an obligation, or making a promise or threat. An epistemic one is concerned with matters of knowledge or belief and with the possibility of something being true. Lastly, dynamic modal verbs are related to the volition or ability of the speaker or subject, in other words, some circumstantial possibility involving an individual (Figures 7 and 8 in Appendix A.1). (Palmer, 1986)

Table 3 shows a contingency table for MoVerb. We see that there is no straightforward mapping allowing us to cleanly convert one framework to the other. However, the different distributions of one set of labels within labels of the other framework reveal which categories are similar to each other.

5 MoVerb: Annotated Modal Verb Dataset

We use the eight core modal verbs in our study: *can*, *could*, *may*, *might*, *must*, *will*, *would*, and *should*. *Shall* is also another core modal verb but is excluded from our work since there are too few instances of it in our conversational dataset.[11] Table 4 shows the statistics of our MoVerb dataset.

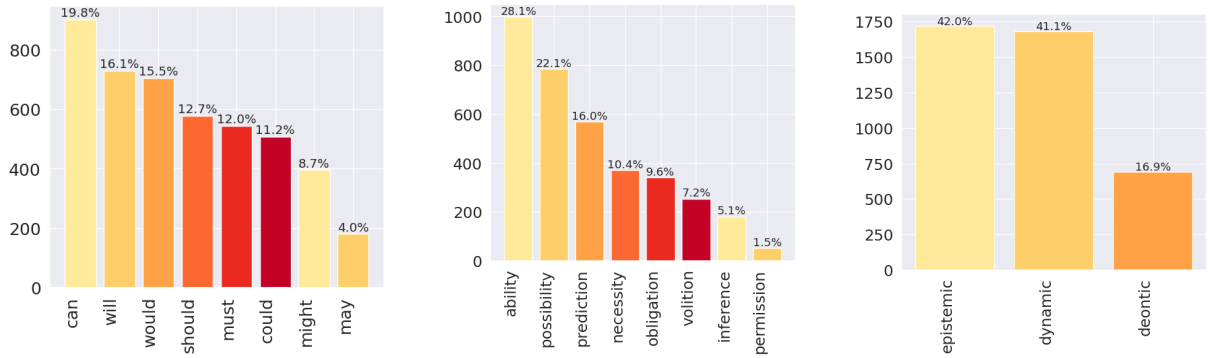
We chose the Empathetic Dialogues dataset (Rashkin et al., 2019) for our annotation task because of its variety of utterances in the conversational domain and wide usage in social dialogue studies. An utterance is defined as a speaker's output in a single turn and can potentially be one or more sentences. We extracted utterances containing only one modal verb as detected using SpaCy's POS tagger and lemmatizer (Honnibal et al., 2020). We focused on utterances containing one modal verb for simplicity, but this excluded very little from the original dataset since only 2.4% of the utterances had more than one modal verb.[12]

We included utterances containing more than one sentence (as long as they used only one modal verb) in order to retain as much context as possible. In this way, we separated out the first 4540 utterances containing single modal verbs, except for **may** and **might**, which we collected and used all of due to scarcity (Table 4 and Figure 1a).

After finalizing which utterances to annotate, we utilized Amazon Mechanical Turk (MTurk) to gather crowd-sourced labels for each modal verb. Three annotations were collected for each of the 4540 utterances, and we assigned final labels based on majority voting (Figures 1b and 1c). We re-

[11] *shall* is more likely to be used in legal contexts (Coates and Leech, 1980), which is outside the scope of this study.

[12] 78.8% had none and 18.8% had one.



(a) Modal verb distribution

(b) Quirk's label distribution.

(c) Palmer's label distribution.

Figure 1: Dataset statistics: (a) Modal verb distribution, (b) Quirk's categories label distribution, and (c) Palmer's categories label distribution. These charts only include utterances that had a majority label.

fer to the annotations for Quirk's categories as MoVerb-Quirk and those for Palmer's categories as MoVerb-Palmer. Our HIT (Human Intelligence Tasks) form, containing definitions and examples for the annotators, is included in Appendix A.1 (Figures 4-10). We limited our MTurk pool to Master workers (high-performing workers) residing in the US with approval rates of $> 98\%$. Each worker was allowed to annotate as many HITs as they wanted and were allowed to submit annotations for both frameworks. They were prevented from participating any further if we saw that their annotations for Quirk's categories seemed random (Appendix A.2). We did not apply the same filter for Palmer's categories because of the less stringent restrictions on which modal verbs each category could be attributed to. However, 95% of our annotators had submitted at least one HIT for each framework, so we were able to apply our criteria to the vast majority of them.

Post-analysis on Annotations Our final annotations revealed some common disagreements (Figures 2, 3 here and Table 11 in Appendix B). In MoVerb-Quirk, annotators seemed to use certain labels interchangeably, as opposed to truly diverging on how the modal verb affected the utterance. For example, in Figure 2, we can see that *inference* and *(logical) necessity* are often confused for the other. Utterances containing sentences like, "You *must* have been so happy" and "You *must* have been so scared" frequently had both *(logical) necessity* and *inference* annotations. Thus, frameworks well-grounded in theory can still be interpreted differently in practice. We see a lack of correlation between sentence length and annotator disagreement (Figure 11 in Appendix B) suggesting that utterance length was not the main or sole cause for this disagreement.

Another common behavior was that annotators seemingly labeled utterances based on what could be inferred. For example, an utterance containing a sentence like "I *may* go to the store today" was often labeled as both *ability* and *possibility*. One could argue that this *may* strongly represents *ability*, since it indicates that the user has the ability to go to the store today. However, one could also claim that the annotator is then labeling what can be inferred from the utterance (if there is a possibility that something would happen, then there exists the ability to make it happen), not necessarily what the modal verb semantically represents.

This behavior can also be observed for MoVerb-Palmer where *epistemic* and *dynamic*, whose definitions overlap with *possibility* and *ability* from Quirk's categories, appear commonly in conflicting annotations (Table 3 and Figure 3). This confusion makes sense when we think of one's ability as the ability to make something possible.

	QUIRK	PALMER
% AGREEMENT	0.58	0.50
KRIPPENDORFF'S α	0.60	0.37

Table 5: Inter-annotator agreement in MoVerb

We see from Table 5 that annotators seemed to struggle more with using Palmer's categories. The percent agreement between the two frameworks was very similar, despite Palmer's categories having significantly fewer labels. We attribute this to the fact that Palmer's categories are more abstract and can thus be less intuitive. The unfamiliar label titles may have also added a layer of complication to the task.

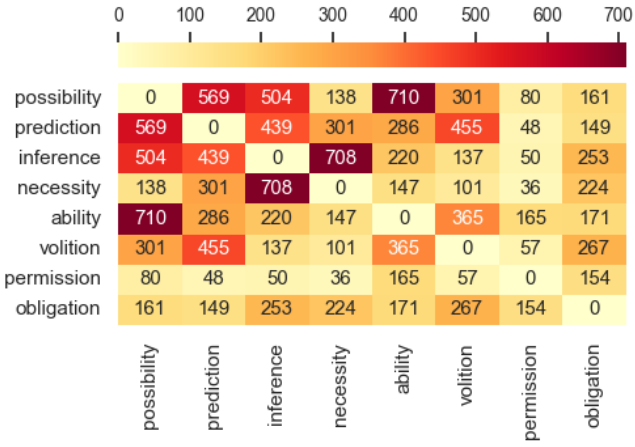


Figure 2: The frequency of disagreeing annotation pairs in MoVerb-Quirk. By disagreement, we mean when two annotators do not choose the same label for some given utterance. Each utterance can have 3 counts of disagreements because there are 3 possible annotation pairs.

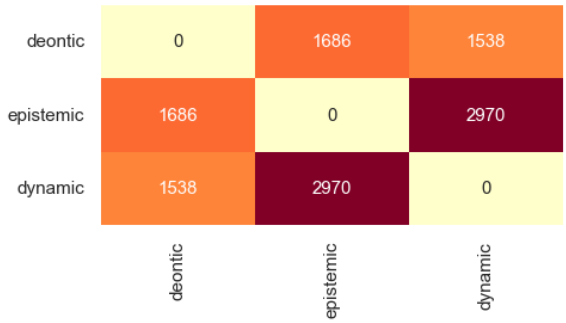


Figure 3: The frequency of disagreeing annotations in MoVerb-Palmer. This uses the same logic as Figure 2

Data Subjectivity We argue that these disagreements highlight the flexibility and ambiguity that have hindered linguists for decades and emphasize the subjectivity of modal verbs. Modal verb annotations highly rely on what the reader interprets as the main takeaway of the modal verb. Quirk's mappings (Table 9 in Appendix B) were not used to limit annotator options in MTurk in order to let annotators select labels with minimal input from us. The added flexibility may have led to lower inter-annotator agreement levels, which is inevitable for subjective annotations. While we provide our scores to showcase inter-annotator agreement as a valuable dataset metric, it should not be solely relied upon to assess the overall quality since it can perceive minority opinions or diversity as undesired noise (Passonneau and Carpenter, 2014; Plank et al., 2014; Aroyo and Welty, 2015; Leonardelli et al., 2021; Basile, 2020).

6 Classification Tasks

We answer the following questions using the collected MoVerb dataset: (1) how well MoVerb can be used to train Transformer models for a modal verb sense prediction task and (2) how transferable that knowledge (trained on the conversational domain) is to other domains, namely the news opinion domain.

Experiment Design In the following experiments, we exclude data where all three annotators disagreed with each other (Appendix C). This is to enable our use of pre-trained models and to share available insights. For our first classification experiment, we split our datasets into cross-validation train-test ratios of 90/10. For the second experiment focusing on transferability, we bring in a third dataset, which we refer to as Ruppenhofer and Rehbein. We use one dataset for training and the other for testing and vice versa. When comparing MoVerb-Palmer and Ruppenhofer and Rehbein, we only consider the overlapping labels since the majority of Ruppenhofer and Rehbein's labels come from Palmer. We conducted this on the setup where both MoVerb-Palmer was the training set and Ruppenhofer and Rehbein was the test set and vice versa. Additionally, since we initially hypothesized that the lack of *will/would* examples in the Ruppenhofer and Rehbein dataset would cause issues, we conducted the same experiment with those modal verbs removed from MoVerb-Palmer (Table 6 and 7).

For all experiments, we ran 10-fold cross-validations and used an early stopping callback that would get triggered once the F1 value stopped increasing by at least 0.01. For learning rates, we tested among $5e-6$, $1e-5$, and $2e-5$, and used the weighted F1 score for evaluation. We used the Pytorch Lightning library to train and evaluate a Transformer model with an Adam epsilon of $1e-8$, and a batch size of 32. Additionally, our trainer used GPU acceleration with a GeForce RTX 3090 using the DistributedDataParallel strategy.

We fine-tuned six Transformer-based models (Vaswani et al., 2017) from Huggingface Transformers (Wolf et al., 2019): $\text{ALBERT}_{\text{base}}$ (Lan et al., 2019), BERT (both base and large) (Devlin et al., 2019), RoBERTa (both base and large) (Liu et al., 2019), and $\text{DistilBERT}_{\text{base}}$ (Sanh et al., 2019). In all runs, the RoBERTa models showed the best test F1 scores (Tables 14 and 15 in Ap-

DATASET	VAL. F1	TEST F1 (BASE)
MoVerb-Quirk	78.98	82.22 (29.9)
MoVerb-Quirk (w/o w^2)	83.56	84.31 (38.3)
MoVerb-Palmer	77.08	78.36 (53.7)
MoVerb-Palmer (w/o w^2)	80.62	80.89 (52.1)
Ruppenhofer and Rehbein	83.31	85.60 (52.0)

Table 6: Best-performing F1 scores averaged over a 10-fold cross-validation. We use w^2 to represent *will/would* and selected the best F1 scores out of various model and learning rate combinations. All scores are from the RoBERTa model due to better performance. For a more complete table, see Table 14. The baseline F1 scores are shown in parentheses, and they highlight the particularly high classifier performance on MoVerb-Quirk.

pendix D). Our loss curves show that our dataset is large enough for these experiments (Figure 12 in Appendix).

Single-Domain Sense Classification From Table 6, we observe that MoVerb can indeed be used to train Transformer-based models (Vaswani et al., 2017) on labeling modal verbs. The table shows that MoVerb-Quirk does better at training models compared to MoVerb-Palmer. We also see that the classifier performs better on Ruppenhofer and Rehbein than on MoVerb-Palmer. This was even after removing *will*s and *would*s, since they were common in our subset of complete disagreements and Ruppenhofer and Rehbein did not annotate those two modal verbs. This greater performance difference may be attributed to the fact that news-related writing tends to be more structured than conversational data and that the Ruppenhofer and Rehbein's dataset contained a higher proportion of *should*s and *could*s, which were less likely to be disagreed upon (Tables 12 and 13 in Appendix B).

Table 8 contains instances where the classifiers predicted incorrectly and with low confidence. Classification of these utterances is especially difficult because of the ambiguity of the modal verbs and room for subjective interpretation. However, this also means the predictions could be used in finding alternative interpretations for some given utterances.

Cross-Domain Transferability We applied the classifiers trained on MoVerb-Palmer to the Ruppenhofer and Rehbein news opinion domain dataset[13] in order to see how our classification model might perform in another domain (Table 7). As mentioned in Section 2, this dataset uses a slightly modified framework, adding three more labels to Palmer's categories. However, we removed them in our experiment since they only made up 3.2% of the dataset we extracted. We also filtered out sentences with more than one modal verb in order to mirror what we use in Empathetic Dialogues (Rashkin et al., 2019).

DATASET	VAL. F1	TEST F1
MoVerb-Palmer → R&R	75.4	61.44
R&R → MoVerb-Palmer	86.5	66.37
MoVerb-Palmer (w/o w^2)→ R&R	80.23	69.74
R&R → MoVerb-Palmer (w/o w^2)	86.5	75.93

Table 7: Observing cross-domain transferability. We use R&R to represent Ruppenhofer and Rehbein and w^2 to represent *will/would*. The dataset to the left of the arrow represents the cross-validation training dataset, while the other is used for evaluation.

We see that our models struggled significantly when the training data and test data came from different sources (Table 7 here and Table 15 in Appendix B). Utterances from a conversational dataset are bound to be different from opinions extracted from news sources due to the nature of their content. We additionally ran the same experiment after removing *will/would* from MoVerb-Palmer to see the extent to which the lack of these two labels affected the F1 scores. The scores rose significantly for both directions although did not reach performance levels observed in single-domain classification. Some difficult examples for cross-domain classification are shown in Table 8 as well.

7 Conclusions and Future Work

Modal verb categorization is a difficult task even for humans, making supervised datasets a vital part of computational analyses. In this study we presented MoVerb, a new modal verb dataset that consists of 4540 conversational utterances with crowd-sourced annotations for the modal verb categories presented by Palmer and Quirk. We show that within MoVerb, annotators struggled less with Quirk's categories. Fewer disagreement relative to the number of labels led to less noise, which translated to better performance on our models, both intra and cross-domain. Additionally, MoVerb-Quirk gave us a more precise study of modal verb patterns due to more specific labels. Therefore, barring cases where there is a specific reason to

[13] http://ruppenhofer.de/pages/Data%20sets.html

DATASET	UTTERANCE	PREDICTION	LABEL
MoVerb-Quirk MoVerb-Palmer	We do not have a fence but I know my dog *will* stay in the yard That stinks! Try not to be jealous though. Something else *will* come your way.	volition (49.36) dynamic (49.87)	prediction (3/3) epistemic (3/3)
Ruppenhofer and Rehbein (R&R)	"A government in which the president controls the Supreme Court, the National Assembly and the Armed Forces *can* not be called a democracy, " Soto charged.	deontic (65.7)	dynamic (N/A)
MoVerb-Palmer → R&R	They are provided with a medical exam upon admission, and their diet ranges from bagels and cream cheese to rice and beans – all eaten with plastic utensils – after which the prisoners *may* clean their teeth with specially shortened brushes.	epistemic (48.95)	deontic (N/A)
R&R→ MoVerb-Palmer	News that big *would* be a shock to anyone! How did you both handle it?	dynamic (49.74)	epistemic (3/3)

Table 8: Difficult examples incorrectly labeled by our RoBERTa-large classifier. The numbers in the parentheses represent the classifier's confidence score for the predictions and the annotator agreement score for the labels. In the first example, we see that the model focuses more on the dog by putting emphasis on its decision (volition) rather than its owner's prediction. In the second example, one could argue that the model focuses more on how one's own actions determine an outcome (dynamic), as opposed to putting more emphasis on plain luck (epistemic). As such, predictions with low confidence levels can help shed light on alternative interpretations.

use Palmer's categories (i.e. expanding another dataset that uses Palmer's categories or comparing work with other studies that use it), we recommend working with Quirk's categories for smoother dataset generation and better downstream task performance. We list limitations of our work in Appendix C.

Our dataset will be available to the public and we hope that it will provide helpful information and insights for other studies as well. Each framework's dataset is split into two subsets: those with a majority label and those with complete disagreement among annotators [14]) (Table 12 in Appendix B). Our fine-tuned classifiers will also be available for those who wish to use them or for combining them with other resources.

This work presents several opportunities for further development. An immediate next step would be to incorporate more modality frameworks into the existing dataset. Potential additional work would be to use the dataset for specific NLP tasks, such as paraphrasing and inference. One way in which modal verbs could be used in inference is to focus on *permission* and *obligation* to see social power dynamics in a text (who seems to be receiving/giving permission more than average or who seems to be controlled by more social obligations). Additionally, one could investigate the annotations with complete disagreements to determine the causes and exhibit high degrees of natural language understanding.

[14]However, this disagreement subset is not used in our experiments.

Ethical Considerations

We paid $1 for 20 annotated sentences on MTurk, which translated to an average hourly wage of $12. This is higher than both the federal and state minimum wage according to the State[15] Department of Labor and Industry. Additionally, recognizing the fact that our HITS were not easy and that annotator blocks can lead to terminated accounts, we utilized qualifications[16] to prevent workers from submitting additional HITS to our project.

This study was reviewed and determined exempt by the Institutional Review Board.

Acknowledgments

We thank Dr. Brian Reese for helpful input during the early stages of the project. We thank Libby Ferland, Petros Karypis, and our anonymous reviewers for their valuable, constructive feedback on the paper. We also extend our gratitude to the Mechanical Turk workers who helped create this dataset.

References

Bas Aarts, April M.S McMahon, and Lars Hinrichs. 2021. *The Handbook of English Linguistics*. Wiley-Blackwell.

[15]https://www.dli.mn.gov/minwage

[16]Qualifications allow us to blacklist workers who did not reach our standards for this particular task, without jeopardizing their account status.

S. Akhtar, V. V. Basile, and V. Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AIIA 2019 – Advances in Artificial Intelligence*, volume 11946, pages 588–603.

L. Aroyo and C. Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36:15–24.

Kathryn Baker, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *DP@AI*IA*.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2002. Longman grammar of spoken and written english. london: Longman, 1999. hardback £69. pp. xii 1,204. isbn 0 582 23725 4. *English Language and Linguistics*, 6(2):379–416.

Jennifer Coates and Geoffrey Leech. 1980. The meanings of the modals in british and american english. In *York Papers in Linguistics*, 8, pages 23–34.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for NLU.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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.

Magali Duran, Adriana Pagano, Amanda Rassi, and Thiago Pardo. 2021. On auxiliary verb in Universal Dependencies: untangling the issue and proposing a systematized annotation strategy. In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 10–21, Sofia, Bulgaria. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303.

Attila Imre. 2017. A logical approach to modal verbs 1. can and could. *Acta Universitatis Sapientiae, Philologica*, 9(2):125–144.

Angelika Kratzer. 1991. Modality. *In Semantics: An international handbook of contemporary research*, pages 639–650.

Angelika Kratzer. 2012. *Modals and Conditionals: New and Revised Perspectives*. Oxford Scholarship Online.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 [cs.CL].

John Lyons. 1977. *Semantics: Volume 1*. Cambridge University Press, London.

Ana Marasović, Mengfei Zhou, Alexis Palmer, and Anette Frank. 2016. Modal sense classification at large: Paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations. *Linguistic Issues in Language Technology*, 14.

Lisa Matthewson and Hubert Truckenbrodt. 2018. Modal flavour/modal force interactions in german: : soll, sollte, muss and müsste.

Dieter Mindt. 2000. *An Empirical Grammar of the English Verb System.* Cornelsen.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Malvina Nissim, Paola Pietrandrea, Andrea Sansò, and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14, Potsdam, Germany. Association for Computational Linguistics.

Jan Nuyts, Pieter Byloo, and Janneke Diepeveen. 2010. On deontic modality, directivity, and mood a case study of dutch "mogen" and "moeten". *Journal of pragmatics*, 42(1):116–34.

F. R. Palmer. 1986. *Mood and Modality*, 1 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

F. R. Palmer. 1990. *Modality and the English modals*. Longman, London.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Paul Portner. 2009. *Modality*. Oxford University Press, Oxford.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. 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 953–965, Online. Association for Computational Linguistics.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? annotating English modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1538–1545, Istanbul, Turkey. European Language Resources Association (ELRA).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

A Experiment Design

A.1 Mechanical Turk Instructions

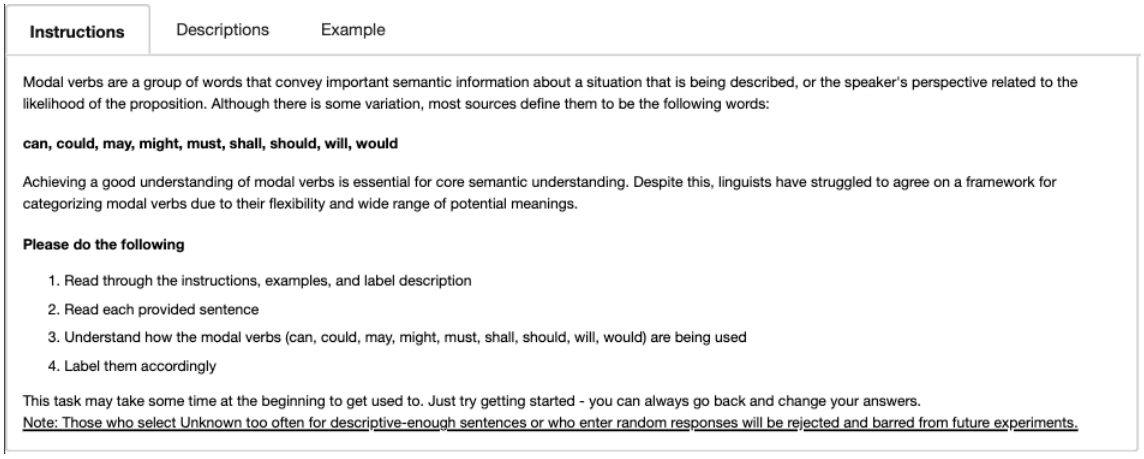


Figure 4: General instructions given to MTurk workers

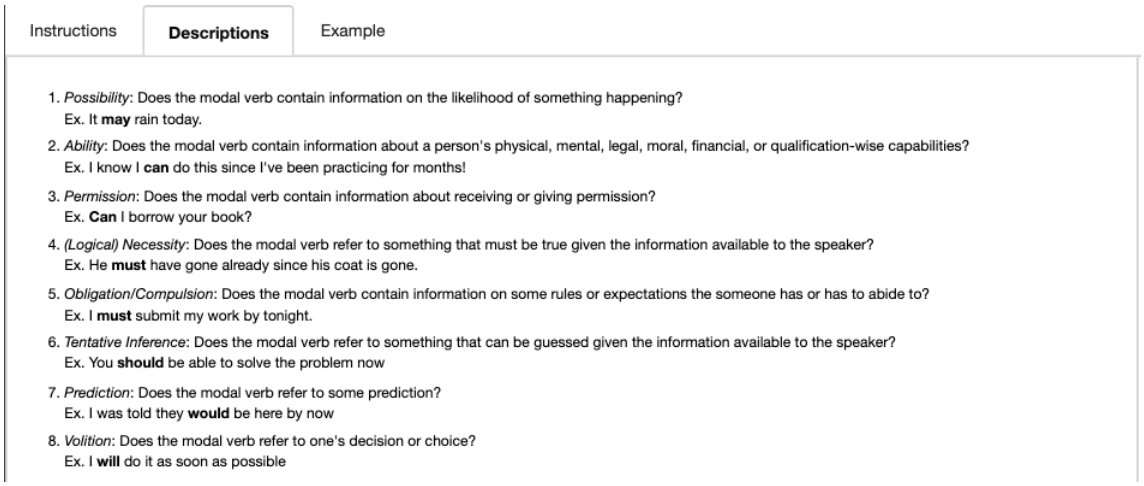


Figure 5: Descriptions given to MTurk workers for Quirk's categories

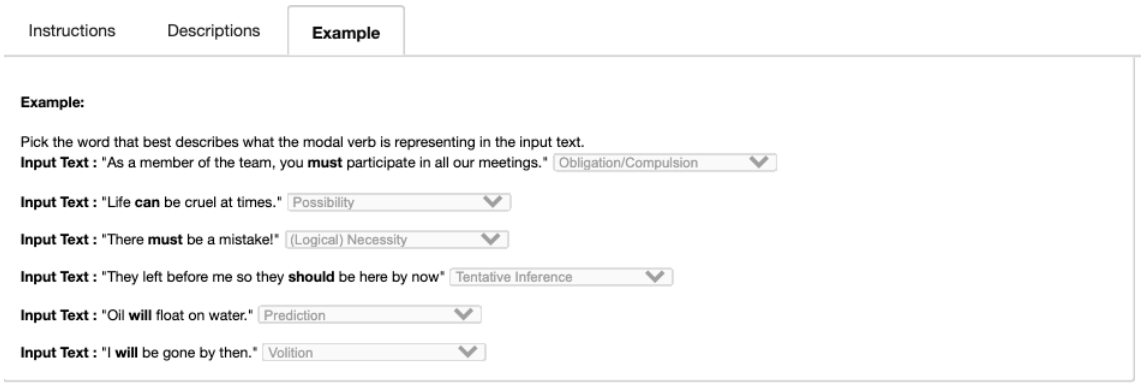


Figure 6: Examples given to MTurk workers for Quirk's categories

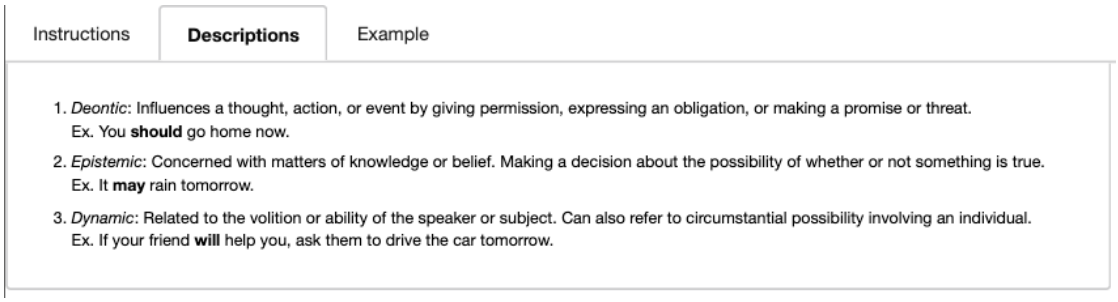


Figure 7: Descriptions given to MTurk workers for Palmer's categories

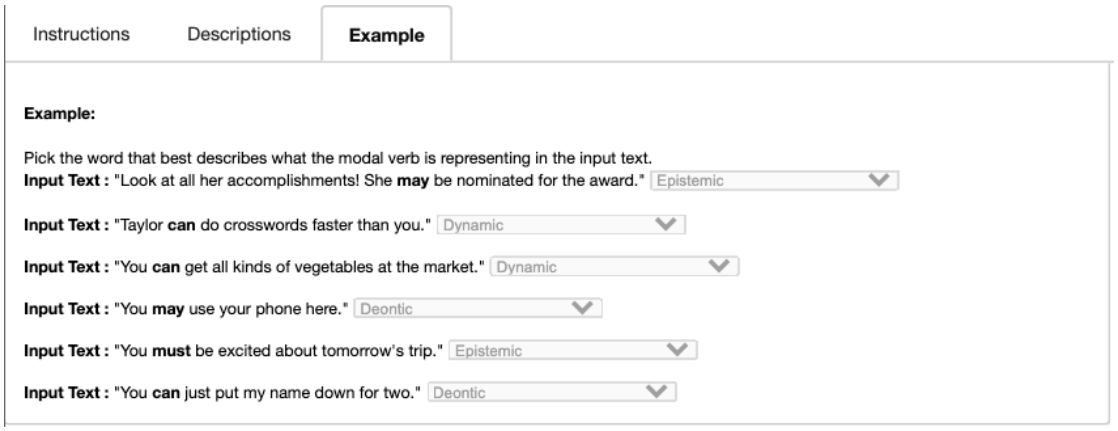


Figure 8: Examples given to MTurk workers for Palmer's categories

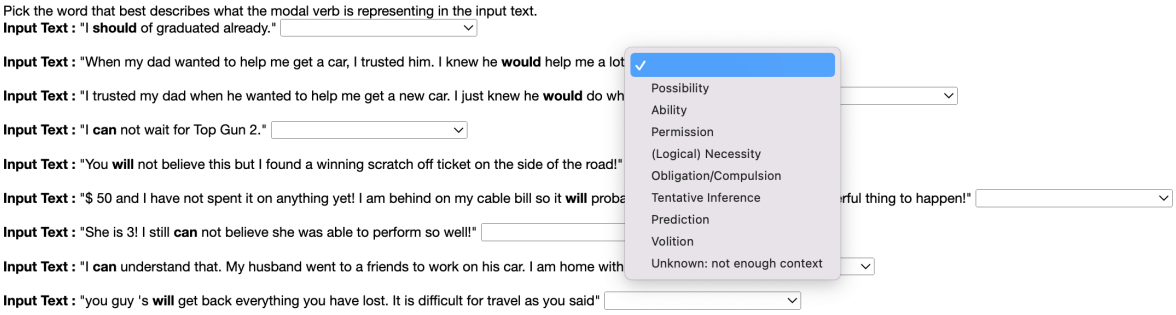


Figure 9: Example sentences to annotate and the corresponding drop-down boxes for Quirk's categories

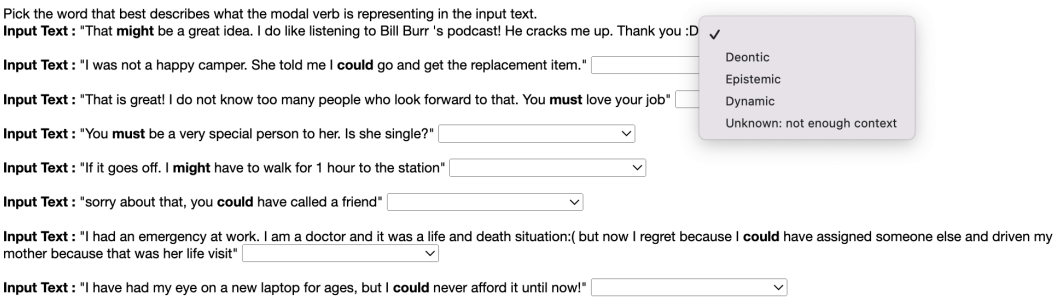


Figure 10: Example sentences to annotate and the corresponding drop-down boxes for Palmer's categories

A.2 Filtering Criteria

Workers were only prevented from working on further HITs when we noticed issues in their annotation quality. The issues were detected based on their frequency of disagreement with others and deviation from Quirk's mappings, which laid out what labels could be assigned to which modal verbs (Table 9). We set the threshold high enough to only filter out the top 1% of whose responses consistently deviated from both their fellow annotators and Quirk's mappings so as to not bias our data. Extreme deviation from both peers and a well-established framework implies more randomness than genuine subjective differences.

	CAN/ COULD	MAY/ MIGHT	MUST	SHOULD	WILL/ WOULD
possibility	o	o	x	x	x
ability	o	x	x	x	x
permission	o	o	x	x	x
necessity	x	x	o	x	x
obligation	x	x	o	o	x
inference	x	x	x	o	x
prediction	x	x	x	x	o
volition	x	x	x	x	o

Table 9: Label to modal verb mapping as defined by Quirk

# ANNOTATIONS	QUIRK	PALMER
< 200	87	83
200 ~400	6	3
400 ~ 600	3	1
600 ~ 800	0	2
800 ~ 1000	1	0
1000 ~ 1200	1	0
1200 ~ 1400	0	1
1400 ~ 1600	0	1
1600 ~ 1800	0	2
$1800 \leq$	3	1

Table 10: Distribution of how many annotations were contributed by each annotator

B Dataset Statistics

COMBINATION	PROPORTION	EXAMPLE UTTERANCES
inference-possibility-prediction	8.00%.	That *should* be fun. Pokemon is a great franchise. I have many of the handheld games.
inference-necessity-prediction	5.16%	The odds *must* be astronomical, almost like winning the lottery.
possibility-prediction-volition	4.45%	Do you mean LeBron James? I was hoping he *would* come to Miami!
ability-inference-possibility	3.54%	Oh no. Were you able to get things sorted out? We live far away from family and I know how hard it *can* be especially when there are health concerns.
ability-possibility-prediction	3.44%	True, just do not like how the world is inching toward a conflict that *could* spill over to a nuclear war.
deontic-dynamic-epistemic	92.63%	I was disappointed by my manager when he told that I *will* probably get my promotion next year(not this year)

Table 11: Top conflicting annotation triplets from MoVerb

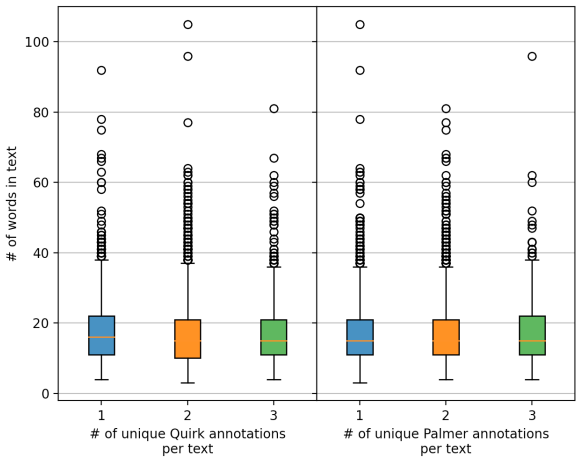


Figure 11: We see no correlation between the number of unique annotations per instance (3 unique annotations would indicate complete disagreement) and the corresponding utterance length. While this is intuitively surprising, it aligns with findings from (Pavlick and Kwiatkowski, 2019).

	QUIRK'S CATEGORIES		
MODAL VERB	AGREEMENT	DISAGREEMENT	TOTAL
will	564	166	730
would	424	280	704
should	406	172	578
may	153	26	179
might	347	48	395
must	431	112	543
could	445	63	508
can	780	121	901
total	3550	988	4538

	PALMER'S CATEGORIES		
MODAL VERB	AGREEMENT	DISAGREEMENT	TOTAL
will	651	79	730
would	592	113	705
should	544	34	578
may	161	18	179
might	358	37	395
must	520	23	543
could	460	48	508
can	805	96	901
total	4091	448	4539

Table 12: Proportion of agreements and disagreements within the dataset. The totals do not add up to 4540 because of "unknown" labels, which we omitted from the table due to low count, but are included in the dataset itself.

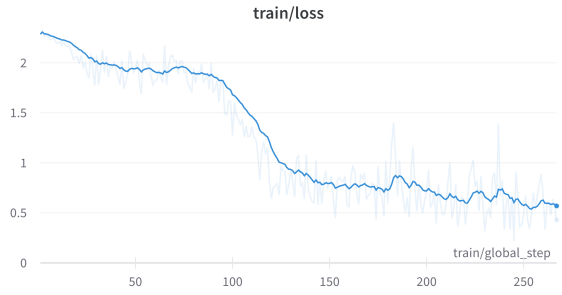
RANK	MoVerb-PALMER		RUPPENHOFER AND REHBEIN	
	MODAL VERB	LABEL	MODAL VERB	LABEL
1	can (19.7%)	epistemic (42.0%)	can (29.5%)	deontic (46.1%)
2	will (15.9%)	dynamic (41.1%)	should (22.4%)	epistemic (27.6%)
3	would (14.5%)	deontic (16.9%)	could (19.7%)	dynamic (26.3%)
4	should (13.3%)	-	must (14.8%)	-
5	must (12.7%)	-	may (8.5%)	-

Table 13: Modal verb and label distribution comparisons between MoVerb and Ruppenhofer and Rehbein (2012). Note that while the modal verb ranking will be the same for both frameworks in MoVerb, we only list a ranking of MoVerb-Palmer in order to compare it with Ruppenhofer and Rehbein (2012).

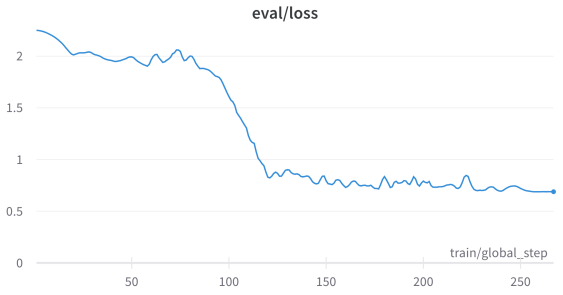
C Limitations

We list several limitations to our work. Firstly, this research does not consider modality in other languages, domains, or frameworks. Our conclusions and insights can only be applied to conversational instances of languages that share the same modal verb morphology as English. Expanding our target text and incorporating more frameworks can thus potentially increase uses for our dataset.
Secondly, our analysis method forces a single label onto each utterance. This is beneficial for training models, but could also mean we are disregarding disagreements that could shed more light onto how people interpret modal verbs. Methods of how to annotate subjective data and handle disagreement have been explored by many (Basile, 2020; Akhtar et al., 2019; Aroyo and Welty, 2015; Davani et al., 2022; Fleisig et al., 2023). We believe our dataset can be used to test these strategies that propose modifications preventing disagreement to be treated as noise. Future work may include allowing annotators to express uncertainty on given labels.
Lastly, since we use crowd-sourced annotations due to resource limitations, we may be missing out on findings that would have been revealed by having more professional or trained annotators. For future work, including input from professional annotators may also allow us to consider frameworks that are more difficult to comprehend in the given time for crowd-sourced workers.

D Classification results

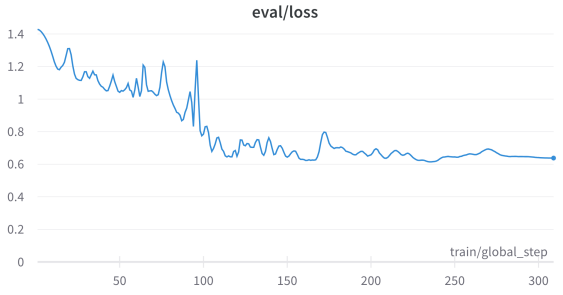


(a) Training loss with 0.8 smoothing for Quirk's subset



(b) Evaluation loss for Quirk's subset

(c) Training loss with 0.8 smoothing for Palmer's subset



(d) Evaluation loss for Palmer's subset

Figure 12: To show our dataset of 4.5K instances is adequate for model training, we present the default loss curve from training a RoBERTa$_{\text{large}}$ model with both Quirk's categories and Palmer's categories.

MODEL	LEARNING RATE	DATASET	VALIDATION F1	TEST F1
$ALBERT_{base}$	5e-6	Quirk	75.49	79.36
$BERT_{base}$	5e-6	Quirk	75.02	77.66
$BERT_{large}$	5e-6	Quirk	77.88	80.56
$RoBERTa_{base}$	5e-6	Quirk	79.21	80.81
$RoBERTa_{large}$	5e-6	Quirk	78.98	82.22
$DistilBERT_{base}$	5e-6	Quirk	78.1	79.19
$ALBERT_{base}$	1e-5	Quirk	69.61	72.67
$BERT_{base}$	1e-5	Quirk	77.84	78.39
$BERT_{large}$	1e-5	Quirk	77.99	80.23
$RoBERTa_{base}$	1e-5	Quirk	78.72	80.53
$RoBERTa_{large}$	1e-5	Quirk	78.63	80.62
$DistilBERT_{base}$	1e-5	Quirk	77.5	78
$ALBERT_{base}$	2e-5	Quirk	70.22	73.18
$BERT_{base}$	2e-5	Quirk	77.74	78.47
$BERT_{large}$	2e-5	Quirk	77.80	79.19
$RoBERTa_{base}$	2e-5	Quirk	78.55	79.88
$RoBERTa_{large}$	2e-5	Quirk	77.42	79.14
$DistilBERT_{base}$	2e-5	Quirk	77.02	77.80
$ALBERT_{base}$	5e-6	Palmer	74.66	75.58
$BERT_{base}$	5e-6	Palmer	76.17	75.49
$BERT_{large}$	5e-6	Palmer	75.22	75.11
$RoBERTa_{base}$	5e-6	Palmer	76.9	77.51
$RoBERTa_{large}$	5e-6	Palmer	77.08	78.36
$DistilBERT_{base}$	5e-6	Palmer	76.37	74.5
$ALBERT_{base}$	1e-5	Palmer	73.63	74.36
$BERT_{base}$	1e-5	Palmer	74.35	74.02
$BERT_{large}$	1e-5	Palmer	74.27	74.68
$RoBERTa_{base}$	1e-5	Palmer	75.94	76.76
$RoBERTa_{large}$	1e-5	Palmer	76.09	76.85
$DistilBERT_{base}$	1e-5	Palmer	74.72	73.6
$ALBERT_{base}$	2e-5	Palmer	74.36	74.79
$BERT_{base}$	2e-5	Palmer	73.66	72.76
$BERT_{large}$	2e-5	Palmer	73.63	74.16
$RoBERTa_{base}$	2e-5	Palmer	75.46	76.57
$RoBERTa_{large}$	2e-5	Palmer	70.54	70.59
$DistilBERT_{base}$	2e-5	Palmer	74.09	72.81

Table 14: F1 scores for fine-tuned models trained using `MoVerb`, averaged over a 10-fold cross-validation.

MODEL	LEARNING RATE	DATASET	VALIDATION F1	TEST F1
$\text{ALBERT}_{\text{base}}$	5e-6	Palmer → R&R	74.26	47.4
$\text{BERT}_{\text{base}}$	5e-6	Palmer → R&R	75.77	42.88
$\text{BERT}_{\text{large}}$	5e-6	Palmer → R&R	75.72	42.29
$\text{RoBERTa}_{\text{base}}$	5e-6	Palmer → R&R	76.89	52.53
$\text{RoBERTa}_{\text{large}}$	5e-6	Palmer → R&R	76.61	54.78
$\text{DistilBERT}_{\text{base}}$	5e-6	Palmer → R&R	75.74	47.71
$\text{ALBERT}_{\text{base}}$	1e-5	Palmer → R&R	71.16	42.09
$\text{BERT}_{\text{base}}$	1e-5	Palmer → R&R	74.8	48.44
$\text{BERT}_{\text{large}}$	1e-5	Palmer → R&R	74.57	50.72
$\text{RoBERTa}_{\text{base}}$	1e-5	Palmer → R&R	75.41	57.99
$\text{RoBERTa}_{\text{large}}$	1e-5	Palmer → R&R	70.47	57.75
$\text{DistilBERT}_{\text{base}}$	1e-5	Palmer → R&R	74.19	54.58
$\text{ALBERT}_{\text{base}}$	2e-5	Palmer → R&R	73.64	52.75
$\text{BERT}_{\text{base}}$	2e-5	Palmer → R&R	74.18	55.72
$\text{BERT}_{\text{large}}$	2e-5	Palmer → R&R	74.29	57.4
$\text{RoBERTa}_{\text{base}}$	2e-5	Palmer → R&R	75.4	61.44
$\text{RoBERTa}_{\text{large}}$	2e-5	Palmer → R&R	70.3	59.1
$\text{DistilBERT}_{\text{base}}$	2e-5	Palmer → R&R	73.7	57.56
$\text{ALBERT}_{\text{base}}$	5e-6	R&R → Palmer	83.41	37.08
$\text{BERT}_{\text{base}}$	5e-6	R&R → Palmer	80.91	56.11
$\text{BERT}_{\text{large}}$	5e-6	R&R → Palmer	81.35	52.35
$\text{RoBERTa}_{\text{base}}$	5e-6	R&R → Palmer	85.76	57.15
$\text{RoBERTa}_{\text{large}}$	5e-6	R&R → Palmer	86.5	66.37
$\text{DistilBERT}_{\text{base}}$	5e-6	R&R → Palmer	82.71	56.36
$\text{ALBERT}_{\text{base}}$	1e-5	R&R → Palmer	81.47	46.08
$\text{BERT}_{\text{base}}$	1e-5	R&R → Palmer	81.82	57.23
$\text{BERT}_{\text{large}}$	1e-5	R&R → Palmer	82.44	53.89
$\text{RoBERTa}_{\text{base}}$	1e-5	R&R → Palmer	85.22	58.2
$\text{RoBERTa}_{\text{large}}$	1e-5	R&R → Palmer	88.07	65.4
$\text{DistilBERT}_{\text{base}}$	1e-5	R&R → Palmer	81.88	55
$\text{ALBERT}_{\text{base}}$	2e-5	R&R → Palmer	80.94	43.96
$\text{BERT}_{\text{base}}$	2e-5	R&R → Palmer	82.74	57.13
$\text{BERT}_{\text{large}}$	2e-5	R&R → Palmer	84.13	58.89
$\text{RoBERTa}_{\text{base}}$	2e-5	R&R → Palmer	84.04	60.71
$\text{RoBERTa}_{\text{large}}$	2e-5	R&R → Palmer	79.61	59.12
$\text{DistilBERT}_{\text{base}}$	2e-5	R&R → Palmer	80.45	57.36

Table 15: Observing cross-domain transferability between Palmer's categories and Ruppenhofer and Rehbein (R&R). We see a clear performance domination of the RoBERTa models.