A Comparative Study of Modal Verb Frameworks with Annotated Datasets

Anonymous ACL submission

Abstract

Modal verbs, such as *can*, *may*, and *must*, are commonly used in our daily communication to convey the speaker's perspective related to the likelihood and/or mode of the proposition (Lyons, 1977; Quirk et al., 1985). They can differ greatly in meaning depending on how they're used and the context of a sentence (e.g. "They *must* help each other out." vs. "They *must* have helped each other out.") Despite their practical importance in areas such as natural language understanding, linguists have yet to agree on a single, prominent framework for the categorization of modal verb senses (Palmer, 1990; Portner, 2009; Kratzer, 2012; Nissim et al., 2013; Torres-Martínez, 2019; Brennan, 1993). 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, we present `MoVerb` which consists of 4.5K annotated sentences from social conversations in Empathetic Dialogues (Rashkin et al., 2019), with each sentence being annotated using two different theoretical frameworks of modal verb senses. We offer insight into the challenges of modal verb ambiguity and suggest modifications when annotating them for downstream NLP tasks. Our dataset will be publicly available upon acceptance.[1]

1 Introduction

Modal verbs (also referred to as modal operators, modals, or modal auxiliaries (Imre, 2017)) convey important semantic information about a situation that is being described or the speaker's perspective related to the likelihood and/or mode of the proposition (Lyons, 1977; Quirk et al., 1985). We will use eight core modal verbs in our study: *can*, *could*, *may*, *might*, *must*, *will*, *would*, and *should*. *Shall* is also another core modal verb, but we exclude it from our study since there are too few instances of it in our conversational dataset. [2]

Because of the widespread use of modal verbs in our daily lives, achieving a good understanding of them is essential for core semantic understanding. In both linguistics and NLP, however, there is no unifying consensus on how to organize these words (Palmer, 1990; Portner, 2009; Kratzer, 2012; Nissim et al., 2013; Torres-Martínez, 2019; Brennan, 1993). One reason for this indeterminacy is the flexibility of their meanings and lack of a straightforward definition (Nuyts et al., 2005). Modal verbs have nuanced meanings, and their interpretation is often determined by who is listening. If a speaker says, "I *can* go to the event today", it can mean that they are capable of going to the event (perhaps they made the time for the party); Alternatively, if the speaker is a minor, the listener may interpret that as having permission from their parents. As such, categorizing modal verbs requires more attention than many other linguistic features, such as tense, where each category is defined by something's place in time and thus can be much more objective.

Although there is some debate as to whether we should focus on modality as a whole since it can be expressed with other vocabulary not limited to modal verbs (Nissim et al., 2013; Pyatkin et al., 2021), we argue that modal verbs by themselves offer enough complexity. There are still downstream NLP tasks that would benefit from better modal verb categorization. Difficulty with modal verb understanding can cause confusion in semantic similarity tasks. Using a RoBERTa Huggingface model pretrained on the Microsoft Research Paraphrase Corpus (MRPC) subset of the General Language Understanding Evaluation (GLUE) dataset, [3] we saw that given some original sentence, the model sometimes would mark all possible senses

[1] `https://anonymized`

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

[3] textattack/roberta-base-MRPC

as correct, even when there was an option that was clearly more likely than the others. 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".[4]

As another example, we generated paraphrases for the Empathetic Dialogues dataset (Rashkin et al., 2019) using the T5 Parrot paraphraser (Damodaran, 2021) in the Huggingface library[5]. This revealed that 1951 out of 2490 (78.35%) paraphrases created for 865 sentences [6] kept their original modal verbs. Thus, being able to correctly identify and paraphrase the sense of a modal verb can greatly increase variety in paraphrasing.

Identifying the meaning and purpose of modal verbs also entails an overall improvement of natural language understanding (NLU). Thus any downstream NLU task could potentially improve from a better understanding of how to identify modal verb senses in NLP. As such, our work can contribute to improving the effectiveness and accuracy of tasks, such as inference drawing, speaker motivation detection, paraphrasing, and question answering.

We present a new dataset, `MoVerb`, containing 4.5K annotated English sentences with their modal verb categories in conversational utterances. We decided to annotate conversational datasets since spoken, casual text is arguably more flexible and nuanced compared to language from other domains, and therefore could reap the most benefits from better modal verb classifications. Additionally, modal verbs are often used with verbs that express one's personal state or stance, such as *admit, imagine*, and *resist* (Krug, 2002), meaning we can utilize them for better speaker intention identification or sentiment analysis as well.

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. We hope this paper will reintroduce a discussion of how we can utilize this family of words to improve results in other areas.

Our main contributions through this paper are:

1. `MoVerb`: An annotated conversational domain dataset containing two types of labels for modal verbs in 4.5K English sentences. We use the majority rule to determine the final label, so our dataset is split into two: the first consisting of utterances with a single final label determined by majority, and the second consisting of utterances with complete disagreement (Tables 1, 4).
2. An in-depth discussion on the challenges of annotating modal verbs and suggestions for alleviating them. We also show that our data can be used in model training and that there is plenty of room for improvement in cross-domain instances. Our fine-tuned classifiers will be released to the public along with our dataset.

Throughout the paper, we use italics for both modal verbs and annotation labels, in order to differentiate them from regular English text. In §2, we discuss the background and novelty of our work compared to prior work on modal verb understanding. In §3, we share our thought process and methodology for collecting this data, and we discuss our findings in §4. Finally, we wrap up the paper with notes on limitations, future work, and ethical considerations in §5 and §6.

2 Related work

There are various linguistic studies about modal verbs and their categorization theories (Quirk et al., 1985; Palmer, 1990; Lyons, 1977; Kratzer, 2012; Mindt, 2000; Aarts et al., 2021). However, despite attempts to reconcile them (Duran et al., 2021), the multitude in variation makes it unclear which theory and framework would work best for various NLP tasks (Figure 9 in Appendix). Having a dataset annotated using multiple modal verb frameworks would help researchers experiment and decide, but that dataset is yet to be built. To the best of our knowledge, there is no English dataset dedicated to the labeling of modal verbs in the conversation domain.

A similar resource available to us is a dataset focusing on subjectivity analysis (Ruppenhofer and Rehbein, 2012). Ruppenhofer and Rehbein (R&R) annotate modal verbs in sentences from opinions

[4] 0.978, 0.732, and 0.988 respectively

[5] prithivida/parrot_paraphraser_on_T5

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

Sentences 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
Sentences 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
And that is something you *should* never do, good on you.	ability	obligation	permission
I work remotely, I wish that you *could* do something like that as well.	ability	permission	possibility

Table 1: Annotation examples from MoVerb for complete agreement and disagreement among three annotators. Note that *necessity* here refers to logical necessity, not social or physical necessities.

and speculations specifically in the news domain (Wiebe et al., 2005). However, news commentators talk about different subjects and use a distinctive variety of vocabulary, implying that a domain shift problem will occur if we attempt to use this dataset for conversational tasks (Li et al., 2019). Krug has shown that modal verbs in different domains, namely conversational and academic, have quite dissimilar distributions. Additionally, R&R do not include *would* and *will* in their annotations, making their dataset unsuitable for analyzing conversational English. *Would* and *will* are 1st and 3rd when we rank modal verbs by their frequencies in spoken English (Mindt, 2000; Krug, 2002).[7]

We should also note that there is a slight difference in our annotation frameworks. R&R create a different schema of their own, building off of work by Baker et al. and Palmer. We do not use the exact same schema since we are more interested in applying traditional linguistic theories. However, we are still able to compare results, since we also heavily use Palmer's work, and 97.57% of the annotations in the R&R dataset are either *dynamic, deontic,* or *epistemic*.

Another similar dataset is an annotated multilingual corpus on modality by Nissim et al. (Nissim et al., 2013). In this work, they use conversational English for a portion of their dataset, but their focus is on modality in general. Thus, they only annotate 32 modal verbs over 7 categories (*will, might, can, may, would, could,* and *should*). Despite this existing dataset on modality as a whole, we felt that a dataset focused on modal verbs was necessary because of the ample complexities of modal verbs even on their own.

[7] As of August 2022, the Corpus of Contemporary American English (COCA) dataset https://www.wordfrequency.info shows *will* and *would* being the 47th and 49th most common words overall. For comparison, *shall* is 1090th.

3 MoVerb: Annotated Modal Verb Dataset

3.1 Frameworks: Quirk vs. Palmer

We use two labeling frameworks in our dataset annotations (Table 2). Although there are other linguists who support the theory behind these categories, we will call them Quirk's Categories and Palmer's Categories for convenience. Proposed categories all build upon each other so it is difficult to credit a single linguist (Table 9 in Appendix). However, both were influential in spreading the frameworks, and we primarily depend on their work for explanations and examples. The labels we use are as follows:

1. 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 each category definition may be inferred by their names, but we include more information on them in Figures 3, 4 in Appendix A.

	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 2: Contingency table showing the frequency distribution between Quirk's and Palmer's categories.

2. Palmer's categories consist of three labels: *deontic*, *epistemic*, and *dynamic*. A deontic modal verb influences a thought, action, 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 whether or not something is 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 5, 6 in Appendix A).

	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 3: Label to modal verb mapping as defined by Quirk

There is a degree of alignment between Quirk's categories and Palmer's categories that makes converting one to the other seem feasible. A number of prominent linguists support the existence and alignment of both categories (Palmer, 1986; Quirk et al., 1985). However, we do not incorporate that mapping into our analysis because these bigger categories often contain an element of linguistic continuum. This means that the labels are on a scale and not in buckets, making it harder to annotate sentences that are in the middle of two labels. In other words, we do not have a quantitative method of determining where one category ends and the other starts (Figure 10 in Appendix).

3.2 Data Collection

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 dialogues. Here, we define utterance as a speaker's output in a single turn - this could be one or more sentences. We extracted utterances containing only one modal verb by detecting them using the SpaCy's POS tagger and lemmetizer (Honnibal and Montani, 2017). We focused on utterances containing one modal verb for simplicity, to avoid modal verb annotations from influencing each other. We conjectured that modal verbs in the same utterances were more likely to be given the same label, even if they meant different things. We included utterances containing more than one sentence (as long as it 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 (Figure 11 in Appendix).

After finalizing our candidate sentences 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 assigned final labels based on majority voting (Table 4). Our HIT (Human Intelligence Tasks) form is included in Appendix A. We limited our MTurk pool to Master workers (high-performing workers) living in the US. Each worker was allowed to annotate as many HITs as they wanted and 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, where he laid out which labels could be assigned to which modal verbs (Table 3). However, we set the threshold high enough so that we would only filter out the top 1% of whose responses consistently deviated from both their follow annotators and Quirk's mappings. We deemed that a high deviation from both implied more randomness than genuine subjective differences. Unfortunately, we could not filter things out for Palmer's categories, since each category has

	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	424	406	153	347	431	445	780	3550
	0.16%	0.12%	0.11%	0.04%	0.1%	0.12%	0.13%	0.22%	
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%)
totals	651	592	544	161	358	520	460	805	4091
	0.16%	0.14%	0.13%	0.04%	0.09%	0.13%	0.11%	0.20%	

Table 4: Contingency table for annotations in adjusted dataset. See Figures 11, 12 and, 13 in Appendix C for a corresponding visual)

less stringent restrictions on which modal verbs can be assigned to them. In other words, there is no clean and unsupervised method of quickly determining data quality.

3.3 Post-analysis on Annotations

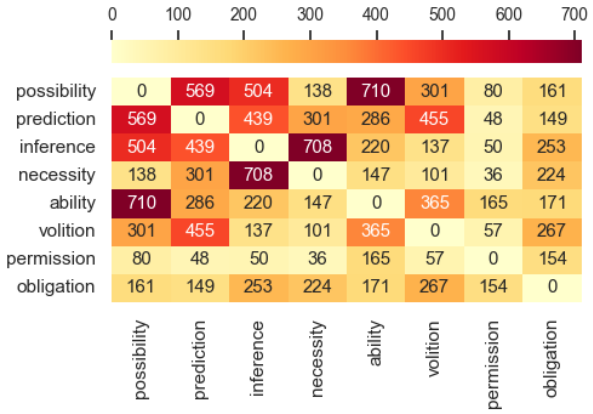

Figure 1: Frequency of disagreement between pairs of Quirk's categories. By disagreement, we mean when two annotators do not choose the same label for some given utterance. Each utterance can have 3 counts of disagreements because there are 3 possible annotation pairs.

Modal verbs are notoriously difficult to categorize (Torres-Martínez, 2019), especially when there is room for interpretation. This can come as a surprise considering the simple labels of Quirk's categories and limited number of options for Palmer's categories. After each HIT, we asked workers to rate the clarity and difficulty of the task, where 10 represented clearest or most difficult. The correlation between these two variables was -0.456 for Quirk's categories and -0.441 for Palmer's categories, showing a significant, but limited negative correlation between the difficulty and clarity of the task.

When reviewing the annotations after our final collection, we noticed there were some common disagreements in the annotations using Quirk's framework (Figure 1). These pairs happened so frequently that our percent agreement value reached only 0.58. However, it seemed that annotators were choosing different labels despite interpreting sentences in similar ways, as opposed to truly diverging on how the modal verb affected the utterance. For example, from Figure 1, we can see that *inference* and *(logical) necessity* are co-occurring in high frequencies. Sentences like "You *must* have been so happy" and "You *must* have been so scared" in the dataset often both had at least one *(logical) necessity* and *inference* annotations each. We can infer from the similarity of how the modal verb is used throughout the dataset that these labels are perhaps being used interchangeably. This illustrates how theoretical frameworks can be interpreted differently in practice.

Another common behavior was that annotators sometimes seemed to label sentences based on what could be inferred. For example, a sentence like "I *may* go to the store today" was often labeled as

both *ability* and *possibility*. One could argue that this *may* represents *ability*, since it indicates that the user has the ability to go to the store today or that the information regarding the speaker's *ability* is most important. However, one could also argue that the annotator is then labelling what can be inferred from the utterance, not necessarily what the modal verb semantically represents.

The percent agreement value, at 0.60, was not high for Palmer's categories either. In Figure 14, we see the frequency of disagreement pairs showing a high occurrence of annotators disagreeing between *dynamic* and *epistemic*. This is not surprising given that *possibility* is often considered *epistemic*, *ability* is *dynamic* (Palmer, 1990), and the two Quirk categories are one of the most common disagreeing pairs (Figure 1).

3.4 Addressing Data Subjectivity

Given that we provide full sets of definitions and multiple examples, we argue that these disagreements highlight the flexibility and ambiguity that have plagued linguists for decades, emphasizing the subjectivity of modal verbs. Using the example above, subjectivity determines whether the listener believes the speaker's *ability* (perhaps they now have time to go to the store) is the main takeaway of the utterance or whether it is the *possibility* that they will go. Quirk's mappings were not used to limit annotator options in the MTurk form since we wanted annotators to select labels on their own with minimal input from us. The added flexibility led to lower inter-annotator agreement levels, but this is inevitable for subjective annotations (Leonardelli et al., 2021; Basile, 2020; Aroyo and Welty, 2015).

3.5 Suggestions for future data collection

Going forward, we propose working with Quirk's categories over Palmer's categories. Despite the significantly fewer labels, the percentage agreement value for Palmer's categories was weak and similar to that of Quirk's categories'. One potential reason is that the framework doesn't categorize modal verbs in a way that is intuitive to lay people. Another reason could be that the unfamiliar category names added a layer of complication to the task.

However, although Quirk's categories work very well theoretically, the amount of overlap and ambiguity that exists can still be challenging when using crowd-sourcing to annotate them. Two simple solutions to this may be to give a clear order of priority (For example, which label should annotators choose when they are stuck between two options?) and to rename the labels so it becomes clearer how one could compare them (perhaps *prediction* could be named *highest possibility*, while *possibility* is renamed to *average possibility*).

4 Evaluation

4.1 Experiment design

After the data preparation, our next focus was to observe how trainable models were using this dataset (Section 4.2) and to see how transferable that knowledge was to other domains, namely the news opinion domain (Section 4.3).

For the first experiment, we split our datasets into train-validation-test ratios of 80-10-10. For the second experiment, since transferability was the focus, we used one dataset for the training data and another for both the validation and test set. We ran this with both a Palmer's categories $\rightarrow$ R&R and R&R $\rightarrow$ Palmer's categories combination. Additionally, since we initially surmised that the lack of *will/would* examples in the R&R dataset would cause issues, we conducted the same experiment with those modal verbs removed from Palmer's categories to observe the effect of not including *will/would* (Table 6).

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. Additionally, we used the Pytorch Lightning library. We use a Pytorch Lightning 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. Our training lasted for approximately 3 hours for each dataset/model combination.

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) (Tables 10, 11). In all runs, the RoBERTa models showed the best test F1 scores (Tables 10, 11).

Dataset	Validation F1	Test F1
Quirk	0.7898	0.8222
Palmer	0.7708	0.7836
R&R	0.8331	0.856

Table 5: Best-performing F1 scores averaged over a 10-fold cross validation. We select the best F1 scores out of various model and learning rate combinations. For a more complete table, see Table 10.

Dataset	Val. F1	Test F1
Palmer→R&R	0.754	0.6144
R&R →Palmer	0.865	0.6637
Palmer (w/o w^2)→R&R	0.8023	0.6974
R&R →Palmer (w/o w^2)	0.865	0.7593

Table 6: Observing cross-domain transferability. We use w^2 to represent *will/would* in the interest of space.

4.2 Single-Domain Classification

From Table 5, we observe that `MoVerb` can indeed be used to train Transformer-based models (Vaswani et al., 2017) on how to label modal verbs. The table shows that Quirk's categories does better at training models compared to Palmer's categories. We know that the framework itself isn't what is causing the disagreement, since R&R uses similar annotations as Palmer's categories but has a higher model performance than Quirk's categories. However, we also wonder whether this raised performance could be attributed to the lack of *will*s and *would*s, which were common in our Disagreement subset. Final statistics for our two subsets can be found in Table 9.

4.3 Cross-Domain Transferability

We also applied the classifiers trained on `MoVerb` (Palmer's categories) to the R&R news opinion domain dataset[8] in order to see how our classification model might perform in another domain (Table 6). 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).

We see that our models struggled significantly when the training data and test data came from different sources (Tables 6, 11). Utterances from a conversational dataset are bound to be different from opinions extracted from news sources due to the nature of their content. Table 7 shows their differences in terms of modal verbs and labels. We additionally ran the same experiment after removing *will/would* from `MoVerb` (Palmer's categories) to see the extent to which the lack of these two labels affected the F1 scores. The scores rose significantly for both directions, but still leave much to be desired (Table 6).

5 Conclusion

We compared two linguistic frameworks by crowdsourcing annotations for 4.5K sentences. Our work shows that within `MoVerb`, Quirk's categories are better suited for supervised NLP tasks due to the greater ease at which annotators seem to label the modal verbs, better performance on the Transformer models, and the more fine-grained labels compared to Palmer's categories.

Additionally, we analyzed patterns found in the annotations and offered potential reasons and solutions to the issues found. Our dataset is available to the public `anonymized` and we hope that it will provide helpful information and insights for other studies as well. Each framework's dataset will be split into two subsets: those with a label majority with at least 2 annotators agreeing with each other (Agreement subset) and those where there were absolutely no agreement among annotators (Disagreement subset[9]) (Table 9). Our fine-tuned classifiers are also released with the dataset for those who need an easy-to-use modal verb intent classifier or find that it can help performance in other tasks when combined with other resources.

6 Future Work

Methods of how to annotate subjective data have been explored by many (Basile, 2020; Akhtar and Patti, 2019; Aroyo and Welty, 2015). These works present how to modify your data or framework, so that disagreement is not treated as noise. We believe out dataset can be used for these shared approaches and believe it would be a worthwhile next step.

Other steps to advance this work would be to use the dataset for specific NLP tasks, such as paraphrasing and bias detection. One way in which

[8]Downloaded from `http://ruppenhofer.de/pages/Data%20sets.html`

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

Rank	MoVerb-Palmer		Ruppenhofer	
	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 7: Modal verb and label distribution comparison

model verbs could be used in bias detection is to focus on *permission* and *obligation* modal verbs to see who seems to be receiving/giving permission more than average or who seems to be controlled by more social obligations. Or perhaps, one could investigate the annotations with complete disagreements and determine what caused those disagreements. Identifying what part of the sentence or context prompted certain annotations and lack of agreement would require high degrees of natural language understanding.

Limitations

We list several limitations to our work. The first is that our data 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 into how people interpret modal verbs. Secondly, this research does not consider modality in other languages, so our conclusions and insights can only be applied to languages that share the same modal verb morphology as English. Lastly, this work only focuses on utterances with single modal verbs. We would need to conduct more studies to determine how generalizable our work is to longer, more complicated sentences. This will be a time-consuming and expensive process; even with one modal verb, the subjectivity of the sentences and fluidity of the modal verbs makes manual inspection crucial to the process.

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[10] Department of Labor and Industry. Additionally, recognizing the fact that our HITS were not easy and that blocks can lead to terminated accounts, we utilized qualifications[11] to prevent workers from submitting additional HITS to our project.

[10]Replaced for anonymity.

References

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

Basile V. Akhtar, S. 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. In *AI Magazine*, volume 36, pages 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*.

Virginia Brennan. 1993. Root and epistemic modal auxiliary verbs.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Magali Sanches Duran, Adriana Silvina Pagano, Amanda Pontes Rassi, and Thiago Alexandre Salgueiro Pardo. 2021. On auxiliary verb in

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

universal dependencies: untangling the issue and proposing a systematized annotation strategy. In *International Conference on Dependency Linguistics - Depling*. ACL.

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

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

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

Manfred Krug. 2002. Douglas biber, stig johansson, geoffrey leech, susan conrad and edward finegan, 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.

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. *ArXiv*, abs/2109.13563.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2019. Domain adaptive text style transfer. In *EMNLP*.

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.

Dieter Mindt. 2000. An empirical grammar of the english verb system.

Malvina Nissim, Paola Pietrandrea, and Andrea Sansòand Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *ACL 2013*.

Jan Nuyts, Pieter Byloo, and Janneke Diepeveen. 2005. On deontic modality, directivity, and mood a case study of dutch mogen and moeten.

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.

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 *ACL*.

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 *ACL*.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? annotating english modal verbs.

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.

Sergio Torres-Martínez. 2019. Taming english modals: How a construction grammar approach helps to understand modal verbs. *English Today*, 35:50–57.

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 Mechanical Turk Instructions

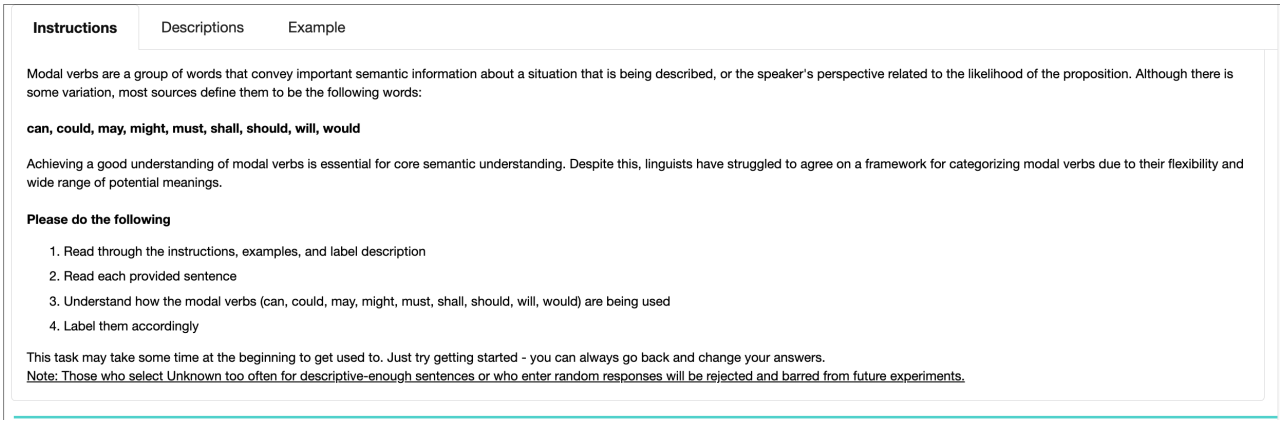

Figure 2: General instructions given to MTurk workers

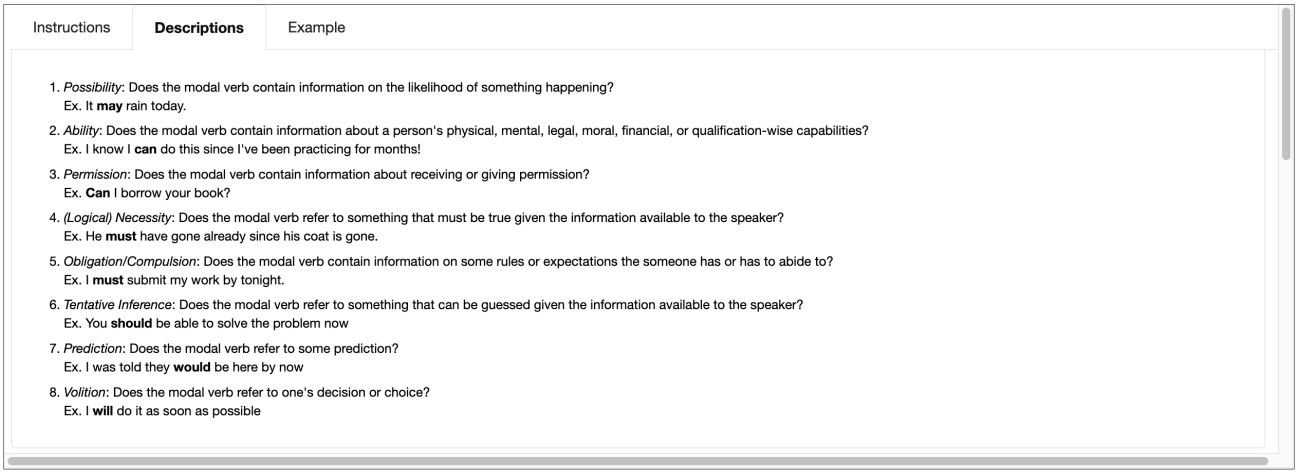

Figure 3: Descriptions given to MTurk workers for Quirk's categories

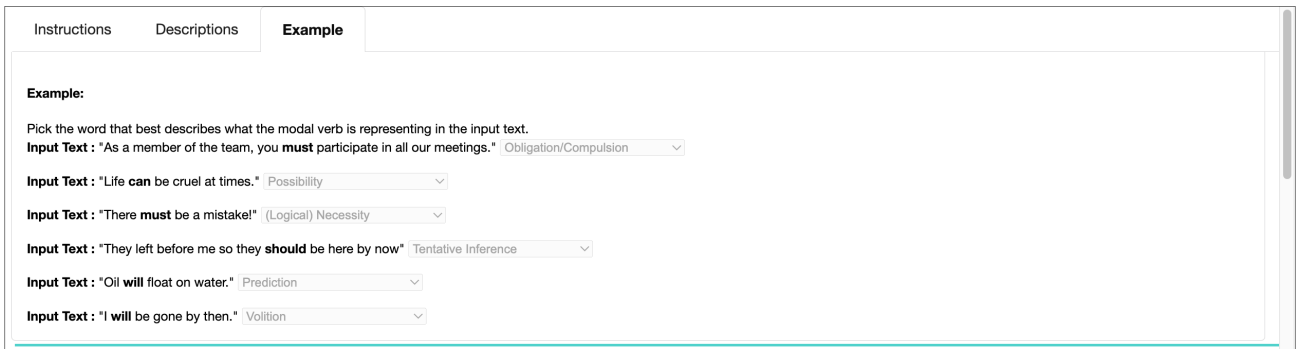

Figure 4: Examples given to MTurk workers for Quirk's categories

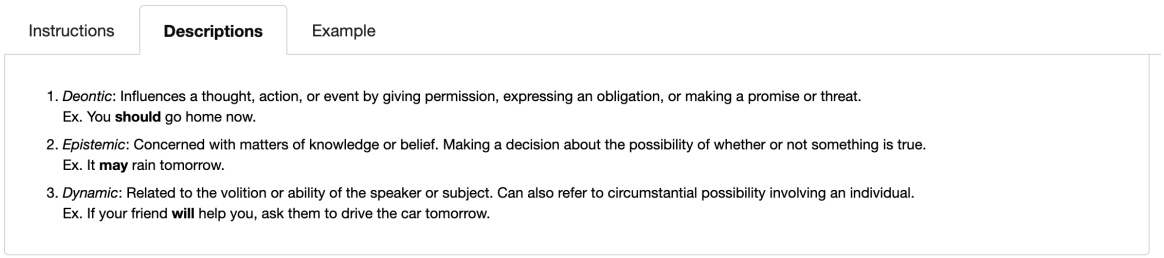

Figure 5: Descriptions given to MTurk workers for Palmer's categories

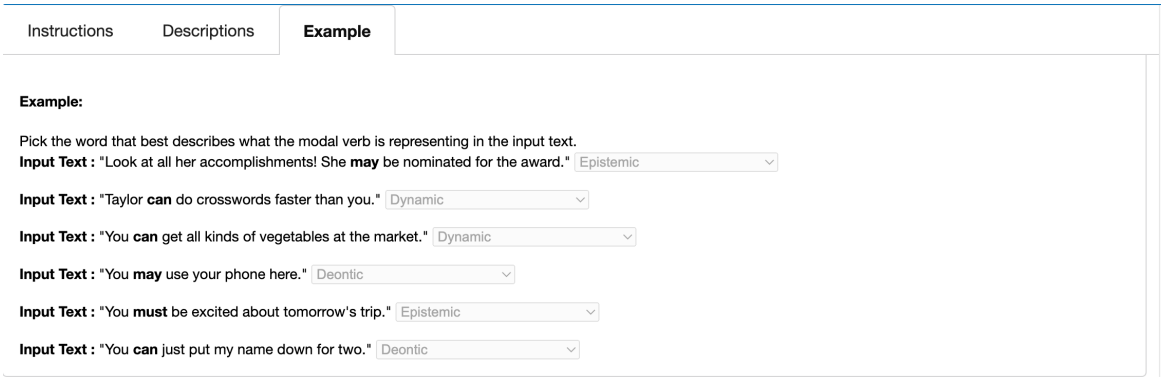

Figure 6: Examples given to MTurk workers for Palmer's categories

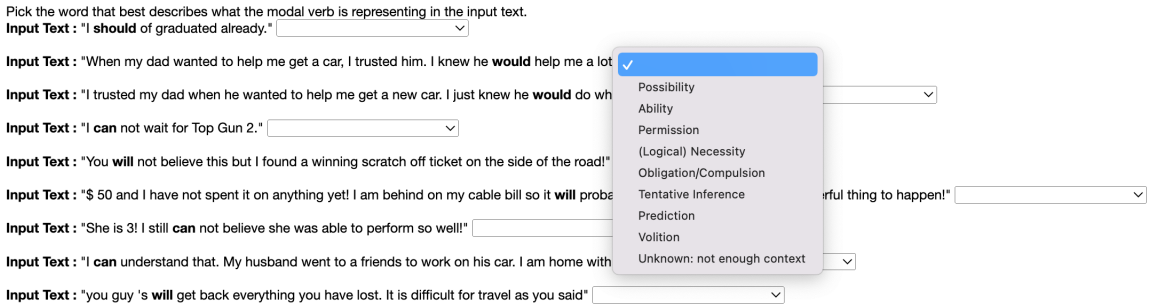

Figure 7: Sentences to annotate and the corresponding drop-down boxes for Quirk's categories

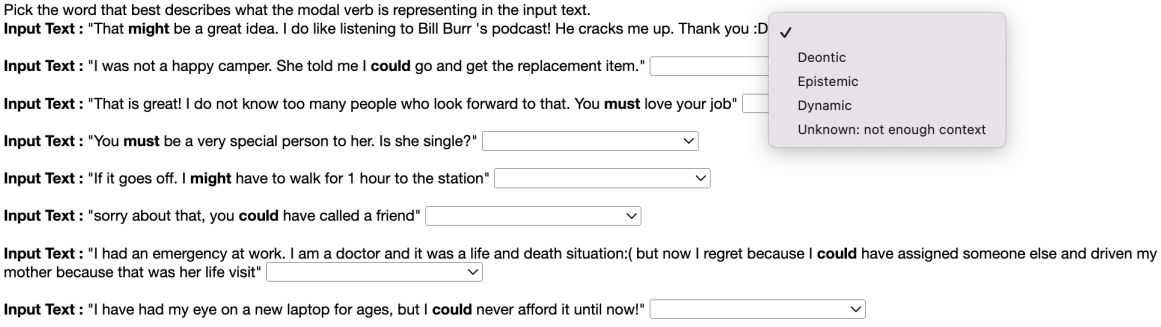

Figure 8: Sentences to annotate and the corresponding drop-down boxes for Palmer's categories

B Modal Verb Categorization

Table 4.1: Semantic classifications for modality

Traditional	Epistemic		Root			
	Epistemic		Deontic		Dynamic	X
My terms	Epistemic		Priority		Dynamic	
		Deontic	Bouletic	Teleological	Volitional	Quantificational
Brennan	Epistemic		Root			
			Deontic		Dynamic	Quantificational
Hacquard	Epistemic	True deontic	Root			X
			Goal-oriented		Ability	

Figure 9: Comparision table from (Portner, 2009), showing how his categorization differs from others. This is a well-created table illustrating how both similar and different linguists can be in labeling modal verbs and their senses.

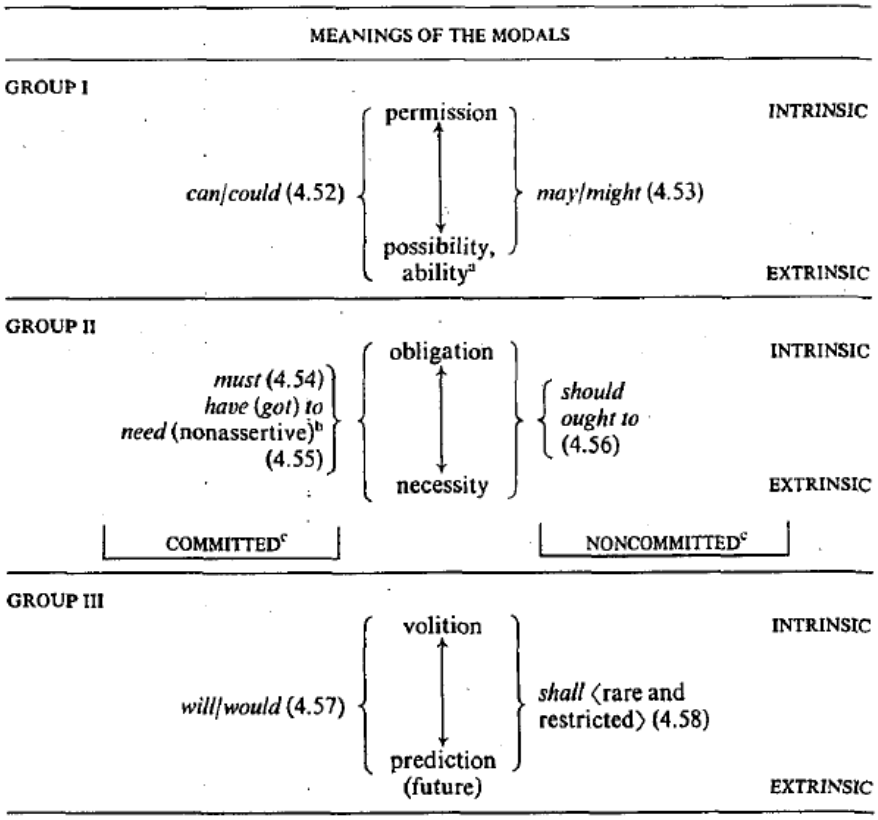

Figure 10: In addition to the challenges of mapping *extrinsic/intrinsic* to *epistemic/deontic/dynamic*, Quirk illustrates the finer categories as ranges within the bigger categories, as opposed to smaller buckets. This raises the risk of incorrectly mapping Quirk's categories to Palmer's categories.

C Dataset Statistics

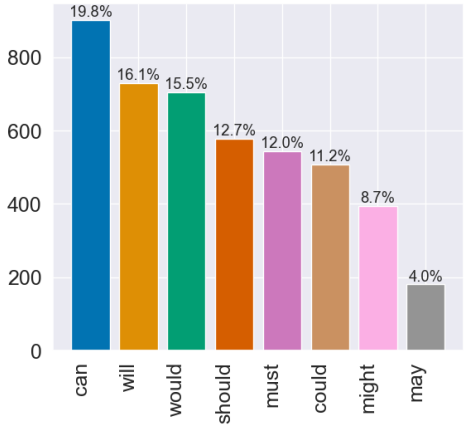

Figure 11: Modal verb distribution

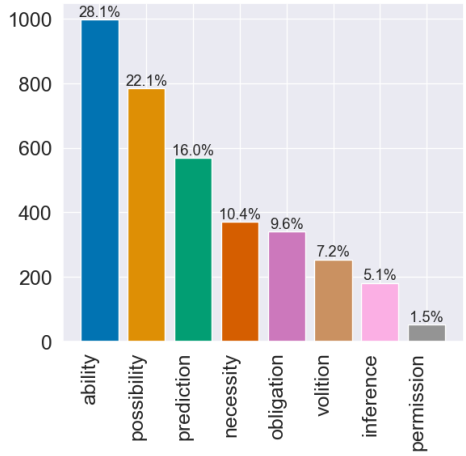

Figure 12: Quirk's categories label distribution. This chart only includes utterances that had a majority label.

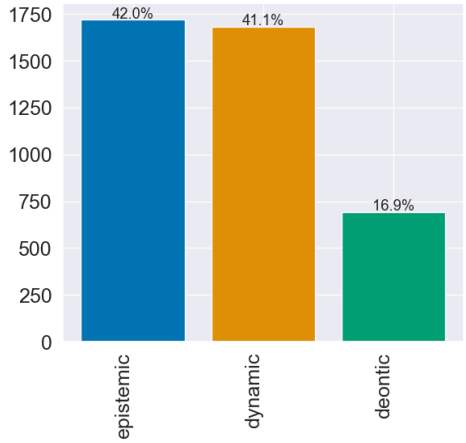

Figure 13: Palmer's categories label distribution. This chart only includes utterances that had a majority label.

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 9: 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.

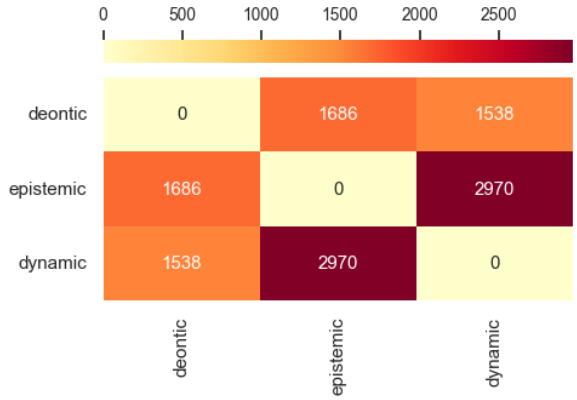

Figure 14: Frequency of disagreement between pairs of annotations in Palmer's categories. This uses the same logic as Figure 1

D Classification results

Model	Learning rate	Dataset	Validation F1	Test F1
$\text{ALBERT}_{\text{base}}$	5e-6	Quirk	0.7549	0.7936
$\text{BERT}_{\text{base}}$	5e-6	Quirk	0.7502	0.7766
$\text{BERT}_{\text{large}}$	5e-6	Quirk	0.7788	0.8056
$\text{RoBERTa}_{\text{base}}$	5e-6	Quirk	0.7921	0.8081
$\text{RoBERTa}_{\text{large}}$	5e-6	Quirk	0.7898	0.8222
$\text{DistilBERT}_{\text{base}}$	5e-6	Quirk	0.781	0.7919
$\text{ALBERT}_{\text{base}}$	1e-5	Quirk	0.6961	0.7267
$\text{BERT}_{\text{base}}$	1e-5	Quirk	0.7784	0.7839
$\text{BERT}_{\text{large}}$	1e-5	Quirk	0.7799	0.8023
$\text{RoBERTa}_{\text{base}}$	1e-5	Quirk	0.7872	0.8053
$\text{RoBERTa}_{\text{large}}$	1e-5	Quirk	0.7863	0.8062
$\text{DistilBERT}_{\text{base}}$	1e-5	Quirk	0.775	0.78
$\text{ALBERT}_{\text{base}}$	2e-5	Quirk	0.7022	0.7318
$\text{BERT}_{\text{base}}$	2e-5	Quirk	0.7774	0.7847
$\text{BERT}_{\text{large}}$	2e-5	Quirk	0.7780	0.7919
$\text{RoBERTa}_{\text{base}}$	2e-5	Quirk	0.7855	0.7988
$\text{RoBERTa}_{\text{large}}$	2e-5	Quirk	0.7742	0.7914
$\text{DistilBERT}_{\text{base}}$	2e-5	Quirk	0.7702	0.7780
$\text{ALBERT}_{\text{base}}$	5e-6	Palmer	0.7466	0.7558
$\text{BERT}_{\text{base}}$	5e-6	Palmer	0.7617	0.7549
$\text{BERT}_{\text{large}}$	5e-6	Palmer	0.7522	0.7511
$\text{RoBERTa}_{\text{base}}$	5e-6	Palmer	0.769	0.7751
$\text{RoBERTa}_{\text{large}}$	5e-6	Palmer	0.7708	0.7836
$\text{DistilBERT}_{\text{base}}$	5e-6	Palmer	0.7637	0.745
$\text{ALBERT}_{\text{base}}$	1e-5	Palmer	0.7363	0.7436
$\text{BERT}_{\text{base}}$	1e-5	Palmer	0.7435	0.7402
$\text{BERT}_{\text{large}}$	1e-5	Palmer	0.7427	0.7468
$\text{RoBERTa}_{\text{base}}$	1e-5	Palmer	0.7594	0.7676
$\text{RoBERTa}_{\text{large}}$	1e-5	Palmer	0.7609	0.7685
$\text{DistilBERT}_{\text{base}}$	1e-5	Palmer	0.7472	0.736
$\text{ALBERT}_{\text{base}}$	2e-5	Palmer	0.7436	0.7479
$\text{BERT}_{\text{base}}$	2e-5	Palmer	0.7366	0.7276
$\text{BERT}_{\text{large}}$	2e-5	Palmer	0.7363	0.7416
$\text{RoBERTa}_{\text{base}}$	2e-5	Palmer	0.7546	0.7657
$\text{RoBERTa}_{\text{large}}$	2e-5	Palmer	0.7054	0.7059
$\text{DistilBERT}_{\text{base}}$	2e-5	Palmer	0.7409	0.7281

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

Model	Learning rate	Dataset	Validation F1	Test F1
$ALBERT_{base}$	5e-6	Palmer → R&R	0.7426	0.474
$BERT_{base}$	5e-6	Palmer → R&R	0.7577	0.4288
$BERT_{large}$	5e-6	Palmer → R&R	0.7572	0.4229
$RoBERTa_{base}$	5e-6	Palmer → R&R	0.7689	0.5253
$RoBERTa_{large}$	5e-6	Palmer → R&R	0.7661	0.5478
$DistilBERT_{base}$	5e-6	Palmer → R&R	0.7574	0.4771
$ALBERT_{base}$	1e-5	Palmer → R&R	0.7116	0.4209
$BERT_{base}$	1e-5	Palmer → R&R	0.748	0.4844
$BERT_{large}$	1e-5	Palmer → R&R	0.7457	0.5072
$RoBERTa_{base}$	1e-5	Palmer → R&R	0.7541	0.5799
$RoBERTa_{large}$	1e-5	Palmer → R&R	0.7047	0.5775
$DistilBERT_{base}$	1e-5	Palmer → R&R	0.7419	0.5458
$ALBERT_{base}$	2e-5	Palmer → R&R	0.7364	0.5275
$BERT_{base}$	2e-5	Palmer → R&R	0.7418	0.5572
$BERT_{large}$	2e-5	Palmer → R&R	0.7429	0.574
$RoBERTa_{base}$	2e-5	Palmer → R&R	0.754	0.6144
$RoBERTa_{large}$	2e-5	Palmer → R&R	0.703	0.591
$DistilBERT_{base}$	2e-5	Palmer → R&R	0.737	0.5756
$ALBERT_{base}$	5e-6	R&R → Palmer	0.8341	0.3708
$BERT_{base}$	5e-6	R&R → Palmer	0.8091	0.5611
$BERT_{large}$	5e-6	R&R → Palmer	0.8135	0.5235
$RoBERTa_{base}$	5e-6	R&R → Palmer	0.8576	0.5715
$RoBERTa_{large}$	5e-6	R&R → Palmer	0.865	0.6637
$DistilBERT_{base}$	5e-6	R&R → Palmer	0.8271	0.5636
$ALBERT_{base}$	1e-5	R&R → Palmer	0.8147	0.4608
$BERT_{base}$	1e-5	R&R → Palmer	0.8182	0.5723
$BERT_{large}$	1e-5	R&R → Palmer	0.8244	0.5389
$RoBERTa_{base}$	1e-5	R&R → Palmer	0.8522	0.582
$RoBERTa_{large}$	1e-5	R&R → Palmer	0.8807	0.654
$DistilBERT_{base}$	1e-5	R&R → Palmer	0.8188	0.55
$ALBERT_{base}$	2e-5	R&R → Palmer	0.8094	0.4396
$BERT_{base}$	2e-5	R&R → Palmer	0.8274	0.5713
$BERT_{large}$	2e-5	R&R → Palmer	0.8413	0.5889
$RoBERTa_{base}$	2e-5	R&R → Palmer	0.8404	0.6071
$RoBERTa_{large}$	2e-5	R&R → Palmer	0.7961	0.5912
$DistilBERT_{base}$	2e-5	R&R → Palmer	0.8045	0.5736

Table 11: Observing cross-domain transferability between Palmer's categories and Ruppenhofer and Rehbein (R&R). We see a clear performance domination of the RoBERTa models.