LiLT volume 14, issue 6 August 2016

Selective Annotation of Modal Readings:

Delving into the Difficult Data

Lori Moon, Patricija Kirvaitis, and Noreen Madden,
University of Illinois at Urbana-Champaign

Abstract

Modal auxiliaries have different readings, depending on the context in which they occur (Kratzer, 1981). Several projects have attempted to classify uses of modal auxiliaries in corpora according to their reading using supervised machine learning techniques (e.g., Rubinstein et al., 2013, Ruppenhofer & Rehbein, 2012). In each study, traditional taxonomic labels, such as *'epistemic'* and *'deontic'* are used by human annotators to label instances of modal auxiliaries in a corpus. In order to achieve higher agreement among annotators, results in these previous studies are reported after collapsing some of the initial categories. The results show that human annotators have fairly good agreement on some of the categories, such as whether or not a use is epistemic, but poor agreement on others. They also show that annotators agree more on modals such as *might* than on modals such as *could*.

In this study, we used traditional taxonomic categories on sentences containing modal auxiliary verbs that were randomly extracted from the English Gigaword 4^{th} edition corpus (Parker et al., 2009). The lowest inner-annotator agreement using traditional taxonomic labels occurred with uses of *could*, with raw agreements of $42\% - 48\%$ ($\kappa = 0.196 - 0.259$), compared to *might*, for instance, with raw agreement of 98%. In response to the low numbers, rather than collapsing traditional categories, we tried a new method of classifying uses of *could* with respect to where the reading situates the eventuality being described relative to the speech time. For example, the sentence *'Jess could swim.'*

is about a swimming eventuality in the past leading up to the time of speech, if it is read as being an ability. The sentence is about a swimming eventuality in the future, if it is read as being a statement about a possibility. The classification labels we propose are crucial in separating uses of *could* that have actuality inferences (Bhatt, 1999, Hacquard, 2006) from uses that do not.

For the temporal location of the event described by a use of *could*, using four category labels, we achieved $73\% - 90\%$ raw agreement ($\kappa = 0.614 - 0.744$). Sequence of tense contexts (Abusch, 1997) present a major factor in the difficulty of determining the temporal properties present in uses of *could*. Among three annotators, we achieved raw agreement scores of $89\% - 96\%(\kappa = 0.779 - 0.919\%)$ on identification of sequence of tense contexts. We discuss the role of our findings with respect to textual entailment.

1 Introduction

Modal auxiliary verbs such as *can, could, would, should, might, must,* and *may* have different readings depending on the context in which they occur. Sentence (1) below provides an illustration:

(1) The president could stop members of congress from expressing their opinions.

On one reading of sentence (1), the author is talking about a past ability the president had to (possibly repeatedly) stop members of congress from expressing their opinions. This reading could occur in a disambiguating context such as that in sentence (2) below:

(2) Back in the 80's, the president could stop members of congress from expressing their opinions.

The reading of *could* is expressing an ability of the president in the past. One can infer that it is possible that there were frequent or at least multiple instances of presidents stopping members of congress from expressing their opinions in the 80's.

On another reading, the author is surmising that there is a possibility of the president stopping members of congress from expressing their opinions in the future relative to the speech time. A disambiguating embedding is given in sentence (3) below:

(3) Due to the problems they are causing, the president could stop members of Congress from expressing their opinions in the future.

For the possibility reading, there is no inference invited that there was even one past instance of the president stopping congress members from expressing their opinions. It seems rather more likely that the president has not stopped them up until the time of speech and that he will stop them in the future.

Readings of modal auxiliary verbs have been of much interest theoretically in linguistics and philosophy due to the nuances of meaning they convey and the difficulties in representing their meaning in formal models. Formal semantic models of modal auxiliary meaning distinguish very fine-grained differences among the taxonomic categories. Those differences will only be alluded to here as they are not the focus of this paper, but they include topics such as which kinds of possible worlds a modal proposition's truth value is evaluated relative to (e.g., Kratzer, 1981,1991; Lewis, 1973), what type of ranking orders the accessible worlds given a modal reading (e.g., Kratzer, 1981; Schulz, 2007; Lassiter, 2011), and the perspective from which the modal is evaluated (e.g., Egan et al., 2005; MacFarlane, 2011; Lasersohn, forthcomming).

Along with the semantic properties of modal auxiliaries, there are interactions among modal auxiliaries, tense, grammatical aspect, lexical aspect, and temporal phrases. These interactions have also been the topic of many studies (e.g., Condoravdi, 2002; Iatridou, 2000; Bhatt, 1999; Bhatt & Pancheva, 2006; Hacquard, 2006).

The fact that modal auxiliary verbs have multiple readings poses several problems for applied areas of research, such as textual entailment and text interpretation. When working on an applied task, however, there must be a balance between accurate representation and broad coverage. Therefore, not everything about modal auxiliary meaning that is of interest can be represented at once. Rather, it is important to focus on the parts of modal auxiliary meaning that most directly impact an automated learner.

One major issue in representing modal auxiliary meaning is regarding the status of the event that is being described by the modality. The event, such as the 'stop-members-of-congress-from-expressing-their-opinions' event referred to in the example sentences above, can be an event that the speaker knows or believes to have happened in the actual world or an event that the speaker predicts in the future or, in the case of counterfactuals, considers not to have happened in the actual world in the past.

In automated tasks, such as question and answering systems, it is important to distinguish among modal auxiliary readings. Given a text that includes sentence (1), an automated system would need to provide a different answer to questions like *'What did the president do?'* or

'Did the president ever stop members of congress from expressing their opinions?' based on the reading of the modal auxiliary verb.

When working on problems in Natural Language Processing (NLP) that require considerable semantic knowledge, it is common to use supervised learning methods, either by having human annotators label a sub-set of the data as training data or by carefully choosing seed set data that reflect patterns typical of the desired classes. When using human annotators, it is important to ensure that what the human annotators are labeling with class labels accurately reflects the concept that the automated system is intended to learn.

Many decisions are involved in word-sense disambiguation tasks. The noun *'bat'*, for example, involves two distinct classes: The flying mammal called a *'bat'* and the instrument for hitting balls called a *'bat'*. The words occur in different positions with respect to their semantic roles: The mammal is likely to be the subject or object of a verb, but the instrument is likely to occur in prepositional phrases. The types of words in the text with each use differs. When the text has words like *'baseball'* and *'stadium'*, the instrument is more likely than when the text contains words like *'cave'*.

The words sense disambiguation problem posed by modal auxiliary verbs, however, poses additional challenges. Modal auxiliary verbs are flexible regarding with which verbs they can occur. Just as one *could* go to a baseball game, one *could* go to a cave. The readings are not limited by their position in the sentence. Two distinct readings can occur in identical strings, as sentence (1) above illustrates.

This paper is about the method we designed to train annotators to label the modal auxiliary *could* according to the temporal properties of the sentence in which it occurs. We focused on *could* for two reasons. The first reason is because results of inter-annotator agreement (IAA) in previous studies show lower agreement on the reading of examples with *could* than on other modal auxiliaries. The second reason is because *could* is the modal auxiliary verb which is most commonly associated with readings allowing actuality inferences.

Section 2 reviews previous studies that led us to isolate uses of *could* from other modal auxiliaries, providing results of our preliminary study. Section 3 presents an overview of the current study. In Section 4, our methods are presented. Section 5 presents the results of our annotation project. Section 6 discusses what the results mean and describes directions for future research.

2 Previous Work on Modal Auxiliary Sense Disambiguation

The distribution of modal auxiliary verbs in corpora has been studied for various applications including corpus linguistics (Coates, 1983), natural language processing (Baker et al., 2012, Ruppenhofer & Rehbein, 2012, Rubinstein et al., 2013), and English language education (Romer, 2004).

Based on previous studies, it is clear that, although each modal auxiliary has multiple readings, not all modal auxiliaries present the same degree of challenges in terms of word-sense disambiguation (WSD) tasks.

For example, the modal auxiliary *might* can be read as expressing a possibility, as in sentence (4) below:

(4) There might be storms today.

In some dialects of English, the modal *might* can also be used to request permission as in sentence (5) below:

(5) Might I have another biscuit?

Although the modal auxiliary *might* has at least two senses, one of its senses is disproportionately more common than the others. In the distributional corpus study by Romer (2004), the modal auxiliary *might* is found to have the possibility sense in 95% of the data observed. Similarly, as shown in the chart in 1, the modal auxiliary *may* has a possibility sense in 83% of the samples she labeled.

In contrast to the modal auxiliaries *might* and *may*, Romer (2004) shows the modal auxiliaries *can*, *could*, and *would* to display a more even distribution of senses. For example, *could* has an *ability* reading in 34% of the labeled data and a *possibility* reading in 41.5% of the data, other significant readings include *hypothetical* (14.5%) and 6.5% of the cases are classified as unclear.

Coates (1983) hand-labeled sentences containing modal auxiliary verbs found in both spoken and written corpora. She presents a breakdown of modal auxiliaries by sense, illustrated in the histogram in Figure 2 below. Modal auxiliaries are positioned relative to each other in the chart given their overall frequency in the text combined with a particular sense.

The histogram in Figure 2 shows that the most frequent modal auxiliary in the text is a *'prediction'* reading of *will*. In order to see, in the histogram, how one reading of a particular modal auxiliary compares to another, it is necessary to find the different instances of the modal

	ability	possibility	permission	hypothet. meaning	prediction	volition	obligation/ advice	inference/ deduction	unclear
can	36%	31.5%	23.5%						9%
could	34%	41.5%	3.5%	14.5%					6.5%
may		83%	13%						4%
might		95%	3.5%						1.5%
will					87.5%	7.75%			4.75%
would				28.5%	50.5%	15.5%			5.5%
shall					31%	65%			4%
should				30%			62.5%		7.5%
ought to				16%			79%		5%
must							52%	39%	9%

FIGURE 1 Romer's 2004 distributional data from the BNC spoken corpus has semantic category labels that reflect traditional labels for modal auxiliary readings. Although we expect North American English written newswire to be considerably different than spoken British English, it provides generalizations that were born out in other studies. (Figure from Romer, 2004).

auxiliary in the histogram. For example, the modal auxiliary *can* with a *'root possibility'* reading is the third most frequent modal auxiliary (by reading). With an *'ability'* reading, the modal auxiliary *can* is the sixth most frequent with less than half as many instances as the *'possibility'* reading. In contrast, the modal auxiliary *might* with an *'epistemic possibility'* reading is the 11^{th} most frequent modal, but no other readings for the modal occur in the chart.[1]

There is additional evidence in previous WSD tasks for the difficulty of *could* compared to other modal auxiliaries. In Ruppenhofer & Rehbein (2012), the authors present the results of their inter-annotator agreement by modal auxiliary. Their table is reproduced in Table 1 below.

The authors combined *could* and *can* into a single category. Their annotation of *can* and *could* is higher than the baseline of the most frequent reading, which they report as 52.17 (2012, page 1545). But their results show that the IAA still is significantly lower for *can* and *could* than for other modal auxiliaries in terms of classification results.

[1]Coates (1983) uses very fine-grained categories for modal auxiliary readings, perhaps too fine-grained for the current WSD tasks. However, the justification for each category is well-presented and provides interesting insights into modal auxiliary readings.

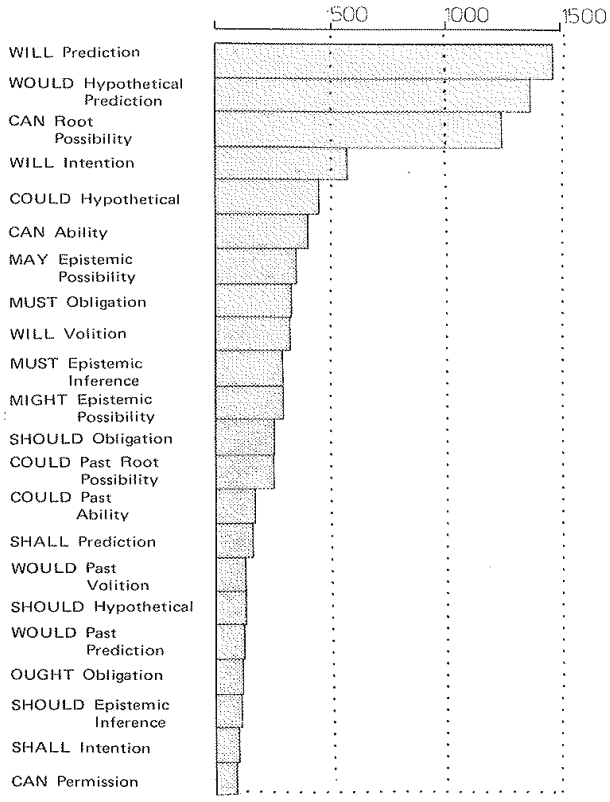

Figure 3.1 Histogram to Show Relative Frequencies of Modal Meanings

FIGURE 2 Distribution of Modal Auxiliaries. This figures shows the distribution of modal auxiliaries broken down by their classification label. Figure from Coates 1983.

	item	κ	%-agreement
may, might	195	0.621	0.89
must	183	0.848	0.98
shall, should	182	0.602	0.96
can	598	0.614	0.77

TABLE 1 Table reproduced from Ruppenhofer & Rebein (2012, page 1544)

The report in Rubinstein et al. (2013) is not broken down by modal auxiliary but rather by reading. All modal auxiliaries are grouped into a single report. Therefore, it is not possible to directly compare our results with *could* to their results without additional information on their classification results, broken down by modal auxiliary. The authors report that they achieved higher results on IAA upon collapsing their semantic categories, but how the collapse alters the results specifically for *could* is not reported.

Baker et al. (2012) provide the results for a single annotator who tagged not only modal auxiliaries but also non-modal auxiliary verbs that were determined by the annotator to be in the scope of the modal and negation. The tags are reported to be 92% correct, however, it is not clear if the annotator accepted one or more than one label for each sentence in determining whether or not the label was correct.

Classification of uses of modal auxiliaries according to semantic taxonomies has been a difficult problem (Baker et al., 2012; Rupenhofer & Rehbein, 2012; Rubinstein et al., 2013). High inter-annotator agreement and automated results have been reported by authors who use only the classes of *'epistemic'* and *'non-epistemic'* (or *'root'*) (Hacquard & Wellwood (2012) for only *must* and *have to*, Rubenstein et al. (2013) after collapsing more fine-grained categories).

2.1 Overview of Testing Traditional Taxonomic Labels

Based on the literature review on the distribution of modal auxiliaries, and based on the results of previous WSD studies on modal auxiliaries, we expected some modals to be more challenging than others for annotators to label. We knew that we would be using a different corpus than the previous studies used,[2] so we wanted to get an idea for how the distributional patterns in our corpus compared to other corpora. We used traditional taxonomic labels such as *'possibility'* and *'ability'* and

[2]There is not a standard corpus across previous studies. We sought the one used in Coates (1983), since it was a well-designed corpus, combining various genres, however, it is not currently available due to the limitations in electronic storage at the time of the publication (p.c. with the author).

tested two hypotheses in a preliminary study: (1) that uses of *can*, *could*, and *would* are more difficult to classify than other modal auxiliaries, predicting lower IAA, and (2) that our results would be comparable to those of previous studies when *can* and *could* are classified according to traditional taxonomic labels.

In order to test the first hypothesis, the modal auxiliaries *should, ought to, may, might, must, have to, will*, and *ought to* were given two senses each according to the labeling used in Romer (2004), as each of these verbs displayed one highly dominant sense in her report and at most one other significant reading (besides the *'unclear'* label). For uses of *can, could,* and *would*, we used the three most common readings listed in Romer (2004) as labels.

In order to test the second hypothesis, we used a finer-grained traditional taxonomy to classify uses of *could*. There is no single standard set of traditional taxonomic senses of *could*, so we borrowed senses from previous studies as they appeared to be useful in examples from our corpus.

We used an annotation schema in which there were eight possible readings of *could*, reflecting traditional ways these modals are interpreted with the addition of the label *'report'* to signify a context in which the reading was unclear due to being embedded under a past-tense verb of saying. The set of labels were *'possibility', 'ability', 'epistemic', 'counterfactual', 'circumstantial', 'permission', 'polite request',* and *'report'*. The set of labels from traditional taxonomies that we used are more fine-grained than those in Romer (2004) but more course-grained than those in Coates (1983). We define them now briefly.

The possibility reading expresses uncertainty about the future.[3]

(6) I'm not certain, but on its current path, the powerful storm could threaten the posh seaside resorts.

(7) Still, a veto could be dangerous for Clinton.

(8) But it could take awhile for the majority to accept gay marriages.

This reading conveys that an event has the potential to come to pass, but the speaker is uncertain or does not have complete knowledge that it will.[4]

The polite reading is used when making a request.

[3]With the exception of simple examples constructed for explanations, the examples in this section are from the English Gigaword 4^{th} edition corpus (Parker et al., 2009)

[4]Some possibility readings are called *'hypothetical'* by some authors when, for instance, they describe something that the author does not feel is necessarily likely

(9) Could I please have a napkin?

Polite requests such as sentence (9) are common in spoken English. Such examples were not present in the newswire data we used, but other polite uses were:

(10) Navarro suggested the pope could undergo an endoscopy examination in order to diagnose his condition, which has prompted at least two bouts of illness - notably at Christmas when the pope had to break off his traditional blessing.

Sentence (10) was taken by annotators to be a polite means of suggesting that the Pope undergo such an examination.

Possibility, hypothetical readings, requesting permission, and polite suggestion uses have in common that they all modalize an eventuality that has not yet occurred at the time of speech.

The epistemic reading is somewhat similar to the possibility reading, but is about non-future states of affairs, as in sentence (11):

(11) The police department is investigating whether something else could have started the explosion.

The counterfactual reading, exemplified in sentence (12), expresses information that is implied to be contrary to known information.

(12) The girl could have been saved, but no one saw her drown.

In sentences (11) and (12), the use of *have* and the perfect form of the main verb aid in choosing the correct label. With the grammatical construction that includes *could have* and the perfect form of the verb, there are only two possible readings, an epistemic reading or a counterfactual reading.

(13) She could have refused the nomination if she had chosen.

In this sentence, the speaker is talking about an event that did not happen, but would have been possible if the *if*-clause had come to pass.

With epistemic readings, the author speculates than an eventuality took place prior to the speech time. With counterfactual readings, the eventuality is also often prior to speech time, but the eventuality did not take place.

The circumstantial reading of *could* expresses that extraneous events have made something possible:

to happen, but that could happen if things in the present were different. This distinction seems to be very subtle in English, so we did not list it as a separate sense.

(14) The report could not be immediately confirmed.

Circumstantial readings were often confused with possibility readings because either can be stated in terms of circumstances. We distinguished them in terms of whether the use of *could* made sense as a past use of *can* or as a use of *could* about a future possibility.

The ability reading expresses a past reading of *can* and is generally regarded as having to do with to personal ability of the subject.

(15) I could swim twenty miles when I was your age.

(16) Cisterns were set up at various parts of the camp, and the women were preparing meals and baking bread with whatever they could scrape together.

The permission reading expresses a higher authority allowing an action or adherence to rules of communities.

(17) Mother said I could go outside and play.

These three readings have very similar grammatical properties and it was difficult to distinguish them in terms of both specification development and annotation.

The remaining reading, labeled *'report'*, involves Sequence of Tense (SoT) contexts (Abusch, 1997) and was used when the reading was unclear to annotators due to the SoT context.

SoT contexts are contexts in which there is a past tense embedding verb, usually a verb of saying. SoT contexts are well-known to result in changing the present tense of direct speech into a past tense in indirect speech (see e.g. Abusch (1997)):

(18) Jess: Syd is sick.

(19) Pat: Jess said that Syd was sick.

At the time when Jess said sentence (18), the state of Syd being sick held. At the time when Pat reports what she said in (19), it is possible that Syd is still sick (but it is not necessary that she be). The past tense form of the main verb, *'was'*, shows grammatical tense agreement with the past tense embedding verb, *'said'*.

A similar scenario occurs when *can* or *will* appears in such a context:

(20) Jess: Syd can come to the party.

(21) Pat: Jess said that Syd could come to the party.

2.2 Methods for Testing Traditional Taxonomic Labels

Four annotators annotated the data. The annotators included the primary researcher and three upper-level undergraduate students.

Samples were taken from the afp_eng_199609 file of the English Gigaword 4^{th} edition corpus (Parker et al., 2009). Documents in the corpus were stripped of all HTML code as well as their document identification numbers. We used a combination of the NLP Toolkit with some added tools to preprocess the files.[5]

We used regular expressions to extract the sentences with modal auxiliaries and created data sets for *must, have to, might, may, will, would, can, could, should,* and *ought to*, as well as negated and contracted forms, such as the *'ll* in *He'll* or the *'wo'* in *won't*. The extracted sentences were randomized using a randomization script in Python. The sentences were stored in files according to the modal they contained.

Data were put into spreadsheets of 25 - 50 sentences, and, for each modal, the annotators were given a list of labels proposed in traditional taxonomies.

Annotators labeled a total of 50 samples of each modal auxiliary and an additional 150 samples of *could*.

2.3 Results of Testing Traditional Taxonomic Labels

The first hypothesis we tested was that the modal auxiliaries *can, could*, and *would* present a more difficult task due to the more distributed nature of their senses.

In order to test this hypothesis, we had annotators label 50 randomized sentences for each modal. If there was one reading in over 80% of the data set, then we set further annotation of that modal aside in order to deal with more difficult cases. We found *should, ought to, may, might, have to, will*, and *ought to* all met these criteria. On a set of 100 randomized samples of *might*, 94% of the uses were given the label *'possibility'*, comparable to Romer's results.

Most of the six epistemic readings were signaled by the presence of *have* and the perfect form of the main verb. We obtained similar results for other modal auxiliaries. In comparison to Romer's corpus, ours was less diverse. Our own pilot work with newswire showed even stronger trends towards a single dominant reading for the modal auxiliaries *might*, *must*, *should*, and *may* than those reported in Coates

[5]The only preprocessing issue that was encountered was regarding tables of comma-separated values of soccer game scores, that are in the English Gigaword 4^{th} edition files. For data such soccer scores, we added the annotator label *corrupt* in order to indicate that it was not possible to annotate the data due to output errors.

(1983).

We expanded the criteria to say that, if there was one reading over 80% of the time given a use of *have* and the perfect main verb form (as opposed to base form) as a feature, then we would exclude those data. Our reasoning was as follows. Some modal auxiliaries such as *must* are almost exclusively epistemic (not deontic) in their uses with *have*. Encoding the presence of *have* and the perfect form of the main verb is possible merely using part of speech tags. Therefore, if such a simple rule could be used to determine the reading, we did not see these as data needing additional annotator effort.

The difficult cases were *can, could*, and *would*. This result was in line with our expectations based on the hand classifications reported in Romer (2004). Of the three difficult cases, we decided to focus on *could* because, in the first set of 50 sentences that we annotated, it proved to be the most challenging.

In order to test the second hypothesis (that, on our corpus, traditional taxonomic labels would get results similar to those reported on other corpora), we annotated instances of *could* according to the taxonomy presented in the previous section.

Our inter-annotator agreement was quite low, only 42-48% (raw) agreement, 0.196 κ - 0.259 κ, as reported in table 2 below. Our results were close to those of Rubinstein et al. (2013) in their report of the version of annotation which did not collapse their proposed labels. An exact comparison is not possible, however, because their results are reported for label accuracy on all modal auxiliaries together, and ours are on label accuracy only for *could*.

	A1 & A2 (n=100)	A2 & A3 (n=100)
% Raw	42	48
κ	0.259	0.196

TABLE 2 Inter-Annotator Agreement on the traditional labels *'ability', 'deontic', 'circumstantial', 'possibility', 'epistemic', 'report', 'permission', 'polite', 'counterfactual'* and *'corrupt'*.

Much of the annotator disagreement was with the use of the *'report'* label, which dealt with sequence of tense contexts. It probably indicates that some annotators found the label more useful than others. In the confusion matrix in Table 3, the major sources of error are in bold font. The matrix shows that 50 errors were due to one annotator choosing the *'report'* label while another annotator choose a different label, such as *'possibility'*.

$n = 200$	P	A	E	CF	C	R	PR	PO	Z
P	54	**5**	1	0	4	**6**	1	0	0
A	**12**	17	1	0	1	**10**	**6**	0	0
E	4	1	4	1	0	0	1	0	2
CF	0	1	1	0	0	0	0	0	0
C	**6**	1	2	0	1	0	1	0	0
R	**19**	**9**	0	0	4	6	0	0	1
PR	1	4	0	0	1	1	7	0	0
PO	0	0	0	0	1	0	0	0	0
Z	2	0	0	0	0	0	0	0	0

TABLE 3 Confusion matrix of traditional senses where P= *'possibility'*, A = *'ability'*, E = *'epistemic'*, CF = *'counterfactual'* C = *'circumstantial'*, R = *'report'*, PR = *'permission'*, PO = *'polite'*, and Z = *'corrupt'*.

Based on the previous literature on corpora studies and word-sense-disambiguation tasks, we anticipated that certain readings are harder to distinguish than others, and we found that hypothesis to hold in our data.

Thus, when we replicated the findings of other annotator tasks using similar traditional taxonomic labels, our results were similar to previous studies, showing *can* and *could* to be the most difficult modals for annotators to agree on. We noticed while annotating the data that past tenses embedded under verbs of saying posed a significant issue in annotation.

3 Overview of the Present Study

In order to improve the classification of uses of *could*, we propose a coarse grouping of senses by the temporal properties associated with them.

We claim that these groupings, in the case of *could*, are more tractable than fine-grained semantic categories such as *'ability'*, *'possibility'*, and *'circumstantial'*. In addition to being more tractable, the temporal categories subsume the finer grained semantic categories.

Some modal auxiliary verbs express past or non-past through the morphology of the modal auxiliary verb and other ones do not express temporal morphology on the modal auxiliary, but appear to express the past through a combination with the perfect form of the main verb.

One type of modal reading that has a past and non-past form is the ability use of *can* and *could*:

(22) Jess can swim a mile.

(23) Jess could swim a mile when she was 45.

One type of modal reading that does not have a past and non-past form morphologically indicated on the modal auxiliary verb is the epistemic use of *might*:

(24) Jess might come to the party tomorrow.

(25) Jess might have come to the party last night.

The modal auxiliary *might* does not change to reflect whether the eventuality is in the past or not. It appears that the past is represented by the use of the perfect form *'have come'* in contrast to *'come'*.

The modal auxiliary verbs that indicate a difference in temporal meaning through the morphology of the modal itself will be called *Paradigm A* modal auxiliaries. Those that do not exhibit a morphological change on the modal auxiliary verb and describe past scenarios via the use of perfect on the main verb will be called *Paradigm B* modal auxiliaries. The modal auxiliary *could* is claimed to have both a Paradigm A sense and a Paradigm B sense.[6]

Paradigm A uses are those in which the present tense of *could* is *can*. The use of *could* is used either to express a past tense use of *can* (such as an ability or circumstantial use) or to match the grammatical past tense of an embedding verb. The semantic readings that fall under Paradigm A are circumstantial, ability, and deontic. We categorized these readings as *Past A* if they were about past eventualities and grammatical past Paradigm A, labeled *GramPast A*, if they were present tense Paradigm A uses displaying SoT effects.

The distinction between these two sub-categories depends on whether the reported speech makes more sense with *can* or *could*. Whenever a use of *could* occurred and was embedded under a past-tense verb, we tested whether the reported speech made more sense with *can* or with *could*, or if the reading made sense with both and was thus ambiguous. If the reading is best with *can*, then the sentence is treated as an instance of grammatical past. An example would be as below:

(26) John said he could make a cake for the party tomorrow.

In sentence (26), it is most likely that John said, *'I can make a cake'* because, in the context of promising to bring something to a party, a firm offer is more felicitous (this is a tendency, not absolute rule). In this case, the past tense of *can* is merely an instance of grammatical

[6] A more complete argument for the paradigmatic division can be found in Moon (forthcoming).

Paradigm B	Present	Past
Inductive	possibility *could*	epistemic *could have*
Metaphysical	hypothetical *could*	counterfactual *could have*

TABLE 4 Categories of Paradigm B readings

tense agreement with the embedding verb. It is a 'fake' past tense (see Iatridou, 2000). If there is no past tense embedding verb, but the sentence seems to be expressing a past use of *can*, then the past Paradigm A *can* label is used, as in the example below:

(27) John could swim when he was a boy.

The semantic category is that of an ability, and it is in the past, so it is a past Paradigm A reading.

The second broad category, Paradigm B, encompasses the uses where past eventualities are referred to by *could have* plus the perfect form of the main verb (e.g., *gone, come, been going*). The subcategories of Paradigm B include: counterfactual, hypothetical, epistemic, and possibility. The semantic categories covered by Paradigm B are shown in table 4.

The readings involve incomplete information on the part of the person making the statement. Possibility is speculation, inductive reasoning about the future based on what knowledge an individual has regarding actual states of affairs. Epistemic reasoning is about states of affairs that are either true or not true in the actual word at the time of utterance, but of which the person making the statement does not know the truth status. The metaphysical readings involve using one's imagination to discuss states of affairs that are not necessarily believed to be true and can be known or believed to be false in the actual world. A hypothetical statement reasons about how things could be now or in the future if some premise held (that the speaker knows or expects not to hold). The counterfactual readings are about how things would have been if a premise held (which the speaker knows does not hold). An example of a hypothetical statement is given in sentence (28):

(28) If I had a car, I could get home easier.

A counterfactual expresses a past state of affairs as in sentence (29):

(29) If I had had a car, then I could have gotten home easier.

In both cases, there is an inference that the speaker does not or did not have a car at the time referenced by the if-clause. The if-clause does not have to be present. It can be contextual, as sentence (30)

below, said in a context where the grandmother is dead and people are thinking of her:

(30) Grandma would have loved seeing the kids.

It should be noted, that because we are using news corpora, the distribution of kinds of readings will be skewed. Note the following example in sentence (31) below:

(31) He was associated with all the major decisions concerning the conduct of the war and could still face war crimes charges from the International Criminal Tribunal in the Hague.

Sentence (31) above is ambiguous. One reading is the reading associated with a historical narrative, in which the author is describing events before speech time. On the narrative reading, the referent of *'he'* was possibly facing war crime charges in the past of the readers present, for example, at a time that the narrative has not yet covered. Historical narrative of this sort is uncommon in news corpora, the source from which these sentences have been pulled.

The second reading, is the most likely reading, given the source of the text. This reading is a non-past reading such that, at the time when the sentence was authored, the possibility of the referent of *'he'* facing war crime charges was something which loomed in the future. [7]

3.1 An Overview of Some Difficult Category Label Distinctions

This section is explains how we went about determining the labels in challenging cases and presents examples. Since the annotators were dealing with data that were not hand-constructed, a number of decisions had to be made regarding how to assign labels. This section concludes with the formulation of three hypotheses that guided the annotation task. The held out data was not discussed among annotators.

[7] It is an interesting fact that the non-past reading merges with the historical narrative reading as the newswire becomes a report of the past for contemporary readers. At this time, nearly twenty years after the writing of the newswire text, it is likely that the possibility of the person referred to being tried for war crimes has been determined. In the annotation task, we considered the reading to be the one that the author was likely to intend at the time of authorship, using simplifying assumptions about our ability to know the author's intent. This problem and similar temporal issues are a subject of study for future research. It is likely to be the case that metadata from text authorship time needs to play an important role in what inferences can be drawn from a text. The notion of author intent is also a complicated concept for guiding the interpretation of linguistic expressions, and characterizing the role it plays would require a more complex model of how meaning is handled by interlocutors than has been developed for this project.

There are two primary label categories that do not fall neatly into either Paradigm A or Paradigm B. The first is *'polite'*, which labels polite uses of *could*. Polite uses are uses in the present, but the speaker uses the past tense to indicate politeness. The sentence below is not past-tense:

(32) You could take some tea, if you would like.

Sentence (32) could be read as employing an ability or permission use of *can* to make an offer in a polite way. If we had to put polite uses into a paradigm, they would fit in Paradigm A, because you can use the present tense form *can*:

(33) You can take some tea, if you would like.

It is not possible to use a past tense with *could have*, as the resulting sentence no longer indicates an offer which holds at the time of utterance:

(34) You could have taken some tea, if you had liked.

The eventualities referred to are not in the past tense with *could*, as other Paradigm A uses are. They are present with past as a grammatical marker, so we include them in the temporally non-past class of Paradigm B.

There are a couple of other difficult cases. The first is where there is not a sufficient context to determine if a reading is from Paradigm A or Paradigm B. As below:

(35) I could quit school at 14.

The sentence above could mean that the person was able to quit school at 14 in the past (past from Paradigm A) or it could mean that the person is not yet 14 and thinking about the future (*'possibility'* from Paradigm B).

The second is where a past-tense embedding verb obscures the meaning making it uncertain whether the reading is Paradigm A or Paradigm B:

(36) John said that he could quit school.

On one reading, John said, *'I can quit school'*, and the reading is a grammatical past form in Paradigm A. On another reading, John said *'I could quit school'*, and the reading is a possibility reading of Paradigm B.

The forms discussed so far appear in table 5.

Sentence (37) illustrates a use of *could* that refers to a past ability.

Modal	Class	Readings
can	Non-Past A	deontic, ability circumstantial
could	Gram-Past A (SoT contexts)	deontic, ability circumstantial
could	Past A	deontic, ability circumstantial
could	Non-Past B	possibility hypothetical
could have	Past B	possibility counterfactual

TABLE 5 Classification Schema

(37) The deep cadence of the religious chant could be heard above the noise of the engine as the jeep-load of Taliban religious militia crossed their own frontline Sunday.

In the past, it was the case that one was able to hear the religious chant above the noise of the engine.

When embedded under a past tense verbs of saying, it can be hard to know whether a reading is a past reading or a grammatical past tense reading of a Paradigm A use of *could*.

(38) Residents said that the deep cadence of the religious chant could be heard above the noise of the engine as the jeep-load of Taliban religious militia crossed their own frontline Sunday.

The use of *'Sunday'* in sentence (38) helps annotators notice that the reading could be either a past Paradigm A reading or a grammatical past reading. This is because it is not clear in the example whether *'Sunday'* describes when the statement was said by residents or when the event being discussed took place.

The Paradigm A past reading means that, at the time when the residents spoke the sentence, they were describing a past state of affairs. In direct quotations, where SoT effects do not occur, the sentence would be as below:

(39) Residents said, "The deep cadence of the religious chant could be heard above the noise of the engine as the jeep-load of Taliban religious militia crossed their own frontline Sunday".

In contrast, if the reading is a grammatical past one, the residents were speaking about a state of affairs concurrent with the time they

were speaking. The direct quotation version would be something like sentence (40) below:

(40) Residents said, "The deep cadence of the religious chant can be heard above the noise of the engine", as the jeep-load of Taliban religious militia crossed their own frontline Sunday.

In examples that have a temporal expression occurring with a direct quotation, the tense of the Paradigm A modal auxiliary must be compatible with the other temporal expressions. If the event of a jeep-load of Taliban religious militia crossing their own frontline occurred before the quoted speech, then it makes sense for the temporal expression *'Sunday'* to be contained in the quotations marks which indicate direct speech. If the event was occurring concurrently with the time of the quotation, then the temporal expression *'Sunday'* makes sense as the reporter's comment on when the utterance was made.

In this example, the various readings seem a bit forced. It seems clearly to be an instance of a past use of a Paradigm A instance of *could*.

Other corpus examples allowing two readings are:

(41) Napatei had ruled that the signatures of the two MPs on a petition requesting the extraordinary session of the 50-member parliament could not be counted.

Since the form is *had ruled*, it is likely that the reading is a past Paradigm A use, and that the use of *could* is a past tense of *can* without the past tense embedding verb. In the past, he ruled saying, "They cannot be counted", however, the ruling is in the past at speech time, so it contributes a past tense without SoT effects.

We recognize that lexical aspect, or Aktionsart (Vendler, 1957), can also affect the temporal location of an eventuality but Aktionsart can be very difficult to identify automatically and we tried to keep our linguistic features tractable. In the data we annotated, we found no cases where issues with Aktionsart would have changed our expected category labels.

Based on our first round of annotation replicating work with traditional taxonomic categories, we came up with three hypotheses to test in our work:

SoT Hypothesis: The difficulty with classifying uses of *could* is partially reducible to the recognition of sequence of tense contexts.

What the *SoT Hypothesis* means is that, if annotators are unaware of SoT effects, they are more likely to disagree on their traditional

taxonomic labels. Therefore, recognizing SoT contexts is an important task in and of itself.

> **Temporally Informed Hypothesis:** Temporally informed categories will lead to more accurate classification than traditional taxonomic labels.

We hypothesized that, if grammatical features of SoT contexts and other sentence-level grammatical features were made a part of the annotation specifications, the IAA would improve. Implicit in this hypothesis was the observation that, without paying attention to grammatical features, annotators tended choose readings that, on a second pass, they did not believe to be accurate.

> **Informative Categories Hypothesis:** The proposed temporally informed category labels assist annotators in their use of traditional taxonomic labels.

The *Informative Categories* hypothesis claims that no information will be lost with respect to traditional labels, but, rather, the temporal labels will inform the traditional labeling and increase IAA on the traditional labels as well.

4 Methods

Three annotators participated over the course of the project: The primary researcher and two upper-level undergraduate students in linguistics.

Our methods for preprocessing data were the same as in the previous study. Samples were taken from the afp_eng_199609 file of the English Gigaword 4^{th} edition corpus (Parker et al., 2009). Documents in the corpus were stripped of all HTML code as well as their document identification numbers. The extracted sentences were randomized using a randomization script in Python.

Preprocessed sentences with *could* extracted from the corpus were put into spreadsheets with drop-down choices among labels. The drop-down choices were new in this round of annotation and helped avoid typographical errors, as well as simplifying the annotation process.

An example of the format that annotators saw is shown in figure 3.

Annotators were given spreadsheets with 25-50 sentences. A total of 600 distinct sentences were annotated and adjudicated for a gold standard annotated corpus with our labeling.

Inter-annotator-agreement was measured on 50 sentences for annotators 1 and 2 and on 26 sentences for annotators 1 and 3 and on 25 sentences for annotators 2 and 3.[8]

[8]The difference in the number of pair-wise measurements was due to the fact

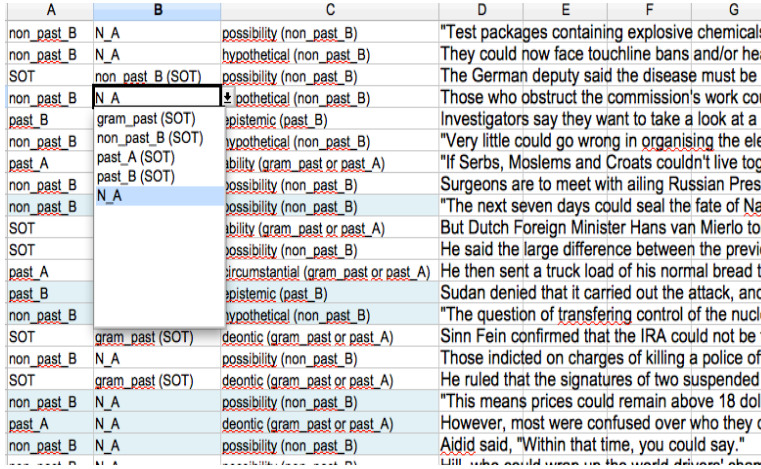

FIGURE 3 An example of the annotation process with drop-down choices shown on one of the three columns

The first column and second column in the annotation software were used together for the final annotation label regarding the temporal category to which a modal reading belonged.

The first column was used to indicate which examples contained a SoT context and, for examples which did not include a SoT context, to annotate the temporal category to which the use belonged. The pull-down menu provided the list of label choices. The label `'SOT'` was used for any examples which contained a SoT context. The other labels were `'gram_past'`, `'non_past_B'`, `'past_A'`, and `'past_B'`.

The second column was used for determining the temporal category of modal auxiliaries which occurred in SoT contexts. The choice `'N_A'`, abbreviating the notion of *'not applicable'* was used for cases in which the first column label was any label besides `'SOT'`. The other temporal categories listed were categories which were compatible with SoT context. These labels were `'gram_past(SOT)'`, `'non_past_B(SOT)'`, `'past_A(SOT)'`, and `'past_B(SOT)'`. The addition of `'(SOT)'` at the end was a reminder to annotators that they were annotating a sentence in which a SoT context was present.

The third column used traditional taxonomic labels like *'ability'* or *'counterfactual'*. We included this column because annotators had used the traditional taxonomic labels in the earlier study and, in certain

that one annotator graduated and took employment abroad near the completion of the task.

instances, they were very confident about the semantic reading.

The third column was only present in this round to facilitate annotators with checking their grammatical and temporal interpretation against possible readings. Although annotators seemed to have stronger intuitions about semantic categories like *'epistemic'* or *'deontic'*, there were very few examples that annotators both found intuitive and assigned the same label.

With each traditional taxonomic label, the temporal label to which it belonged was listed, as shown in the display pull-down menu in figure 3. When annotators viewed the pull-down menu, for instance, to see *'ability'* they would see the choice `'ability (past_A or gram_past)'`. This added an additional check-point to annotation accuracy. If an annotator thought a reading was an ability reading, but had chosen the `'non_past_B'` label in column two, then she was aware that she had mislabeled one column.

Annotators read each sentence and were encouraged to fill out whichever of the three columns was most immediately obvious to them and then proceed to the other columns and check that all were consistent with each other, consistent with the sentence, and consistent with the decision-tree algorithm explained in the next subsection.

4.1 Decision Tree of the Annotation Algorithm

Annotators were asked to check whether or not the instance of *could* occurred with the perfect form of the verb and *have*. If so, the main label was `'past_B'`. In our paradigm descriptions, only past Paradigm B modals occur in the grammatical construction *could have* with a perfect verb form.

Once a reading is determined to be a past Paradigm B reading, the choices in column three are reduced to *'epistemic'* or *'counterfactual'*. The eventuality being described is something situated in the past relative to the speech time. In either case, classifying the sentence as having a past Paradigm B use of *could* places the eventuality being modified by the modal auxiliary in the past relative to the time at which the sentence was uttered.

If the modal auxiliary does not occur with *have* and the perfect form of the main verb, then there is only one other option for the main verb form: *could* occurs with the base form of the main verb.

If this is the case, then annotators consider whether or not it is in a SoT context. If it is not in a SoT context, the annotators label column two as `'N_A'` for the notion of *'not applicable'*.

Next, annotators consider whether the eventuality being described

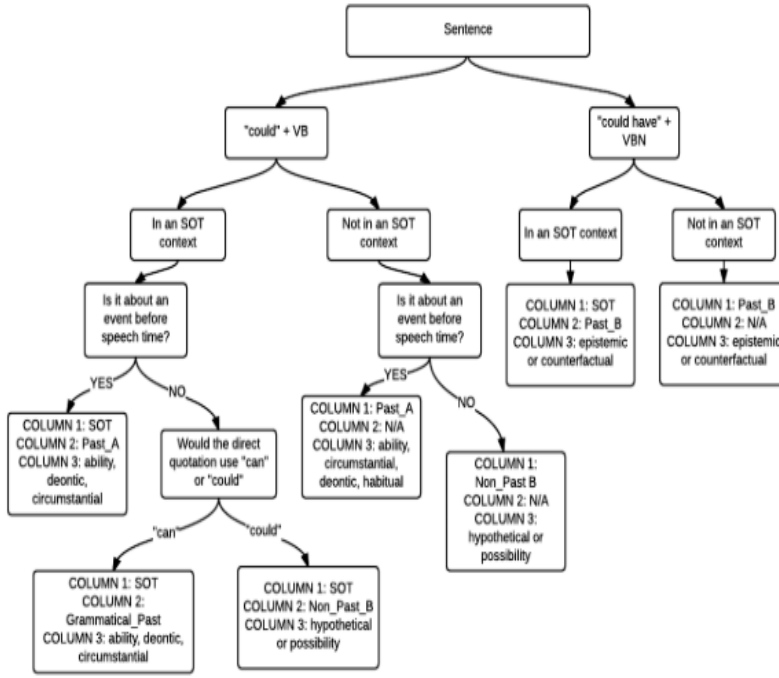

FIGURE 4 Decision tree for annotator guidelines.

is before the speech time or not. If the event is before the speech time, column one is labeled as a past form of Paradigm A. The semantic reading in column three is then an ability, circumstantial, or deontic reading.

If the eventuality is not before the speech time, the column one is labeled as a non-past form of Paradigm B, and column three is labeled either as a possibility or hypothetical reading.

If the use of *could* with the base form of the main verb does occur in a SoT context, then annotators mark column one as being SoT.

If the sentence is about an eventuality that is non-past, the annotators label column two as a past Paradigm A form. Column three can be an ability, deontic, or circumstantial reading.

If the sentence is about an eventuality that is before the speech time, then annotators have to decide between grammatical past Paradigm A readings and non-past Paradigm B readings. If, imagining oneself at speech time, the direct quotation sounds natural with *can*, then column two is labeled as a grammatical past form of Paradigm A. Column three can be an ability, deontic, or circumstantial reading.

If the direct quotation seems more natural with *could*, then column two is labeled as a non-past form of Paradigm B, and column three is labeled as a possibility reading.

The abbreviated method appears in figure 4.

5 Results

This section presents the results of our annotation using the categories that we proposed.

It was hypothesized that SoT contexts play a significant role in the issues that annotators have agreeing on senses of *could*. In order to test the hypothesis that awareness of SoT contexts would improve annotation, we built the recognition of SoT contexts into the annotator specifications and placed a separate column for labeling SoT in the annotation software. We found that, when made aware of SoT contexts, annotators could identify them with high accuracy. The IAA on labeling sequence of tense is shown in table 6 below.

	A2 & A3 (n=25)	A1 & A2 (n=50)	A1 & A3 (n=26)
% Raw	88.8	92	96.1
κ	0.779	0.911	0.919

TABLE 6 Inter-Annotator Agreement between whether the annotators identified the context as a sequence of tense context or not.

We achieved between $0.779\kappa - 0.919\kappa$. We have no comparison with other studies as we are unaware of other SoT labeling projects.

The SoT Hypothesis stated that identification of SoT contexts assists with identification of the major categories.[9]

The SoT Hypothesis and the Temporally Informed Hypothesis were both tested with respect to the temporally-inspired category labels *'Past A', 'Grammatical Past A', 'Non-Past B'*, and *'Past B'*.

We hypothesized that the identification of SoT effects would help annotators identify the temporal category. We also hypothesized, more generally, that, if annotators are made aware of how they construe the temporal placement of the event described in the sentence with the modal auxiliary, they are more likely to agree on a classification.

We tested these hypotheses through our training of annotators and the specifications that prompted them to consider the temporal placement of events when considering a modal auxiliary reading.

We found that the results of annotating with temporally informed categories indicated significantly higher IAA over the use of traditional taxonomic categories. We achieved between 73% and 90% raw agreement (κ of 0.61-0.74) using the temporally informed categories. Our IAA using this method is shown in table 7.

[9]It is important to note that errors in SoT context identification do not necessarily lead to errors in paradigm labels. Annotators who failed to identify SoT contexts could yet label the paradigm form correctly.

	A2 & A3 (n=25)	**A1 & A2 (n=50)**	A1L & A3 (n=26)
% Raw	88	**90**	73
κ	0.744	**0.744**	0.614

TABLE 7 Inter-Annotator Agreement on the four labels *'Past A'*, *'Grammatical A'*, *'Non-Past B'*, and *'Past B'*.

Most of the errors that occurred among annotators were about the difference between *'Non-Past B'* on one hand and *'Past A'* or *'Grammatical Past A'* on the other. A confusion matrix of the decisions is given in table 8 below:

$n = 101$	PastA	Gram-PastA	Non-PastB	PastB
PastA	19	1	**2**	0
Gram-PastA	**3**	5	**3**	0
Non-PastB	**5**	**2**	75	2
PastB	0	0	0	4

TABLE 8 Confusion matrix of *'Past A'*, *'Gram-Past A'*, *'Non-Past B'*, and *'Past B'*.

Overall, requiring explicit identification of SoT contexts in the annotation software and giving annotators specifications that made them consider the temporal and aspectual properties of the verb phrase resulted in improved annotation of the most difficult modal auxiliary data.

Our third hypothesis, the Informative Categories hypothesis predicted that the use of the temporal labels would only have a positive impact on the use of traditional labels.

We tested this hypothesis by including a third column for traditional labels in the annotation software. Although we are proposing a replacement the traditional labeling, we calculated the results on the third column as an extra measure to check that the temporal labels assisted in traditional label identification.

We found that the temporal categories indeed positively affected the assignment of traditional categories. Our results on these data were an improvement over our previous study on different sentences from the same corpus (which was presented in Section 2). We achieved $69\%-80\%$ raw agreement ($\kappa 0.558-0.687$). The results are shown in table 9 below.

Most of the errors with the traditional labels were confusions among possibility readings and ability readings, which are *'Non-Past B'* versus *'Past A'* or *'Grammatical Past A'* errors in the temporal classification. The confusion matrix appears below:

	A1 & A3 (n=26)	**A1 & A2 (n=50)**	A2 & A3 (n=25)
% Raw	69	**84**	84
κ	0.558	**0.619**	0.687

TABLE 9 Inter-Annotator Agreement on the traditional labels *'Ability', 'Deontic', 'Circumstantial', 'Possibility', 'Epistemic',* and *'Counterfactual'.*

$n = 101$	ability	deont	circum	posso	epist	couterf
ability	16	3	0	**9**	0	0
deontic	1	0	0	0	0	0
circumst	2	0	0	0	0	0
possibility	0	1	0	59	0	0
epistemic	0	0	0	2	3	0
counterf	0	0	0	0	1	3

TABLE 10 Confusion matrix of traditional senses *'ability', 'deontic', 'circumstantial', 'possibility', 'epistemic',* and *'counterfactual'.*

The results show an increase in IAA over using traditional categories and support the hypothesis that the temporal categories offer an insightful way to annotate uses of the modal *could* according to the reading with which it is associated.

The next section discusses what the results mean in terms of the hypotheses and the previous literature, applications and future research directions.

6 Discussion

Our scores on the most difficult modal auxiliary data are comparable to the IAA results of other projects, however, the results indicate progress in annotation because the other projects report IAA results for data sets which include modal auxiliaries which have been shown to be easier than *could* for annotators to label. Our SoT Hypothesis is further supported in that, of 20 total disagreements on traditional taxonomic senses across three annotators, 18 of the disagreements involved SoT contexts. Of those disagreements, half involved a disagreement on whether a reading of *could* in a SoT context was stating an ability (either past or concurrent with the speech time) or a possibility. The disagreed-upon sentences allow either interpretation and are only definitively resolvable with a larger text window. The fact that annotators disagreed on data that, upon adjudication, do not have a clear reading, indicates that our specifications have produced a labeling system that reflects genuine ambiguities.

6.1 Comparison with Previous Studies

We do not have a direct comparison using the same type of data and taxonomy. Ruppenhofer & Rehbein (2012) provide the most similar classification that allows for comparison. Their results, as mentioned before, are repeated in Table 11 below:

	item	κ	%-agreement
may, might	195	0.621	0.89
must	183	0.848	0.98
shall, should	182	0.602	0.96
can	598	0.614	0.77

TABLE 11 Table reproduced from Ruppenhofer & Rebein (2012, page 1544)

In Ruppenhofer & Rehbein (2012), uses of *can* and *could* were combined. Annotators classified readings according to three categories: epistemic, dynamic, and deontic.

It is impossible to say without carefully viewing the authors' annotation data how the categories correspond to the ones we used, but it is possible to get an idea from the examples in their paper. The authors give examples of *can* according to each of the three categories, and, later, a list of uses of *could* which they claim has all the readings that *can* has, though some are only possible in shifted tense contexts. They do not label the examples with *could* according to how they intend them to be classified, but most of the six examples are fairly clear.

The paper describes the epistemic label as follows: 'The epistemic use is concerned with the speaker being compelled to come to a particular conclusion given her state of knowledge.' Although this description is a simplification of assessment-sensitivity in epistemic semantics, it is an understandable simplification given the broad scope of the task. They state that *could* is epistemic in the sentence, *'The NHL star hinted he COULD be in the lineup.'*. We would label this sentence as being a non-past Paradigm B use. That use would be further reducible to either a hypothetical or possibility use. We would expect to see a number of temporally situated eventualities in the set of uses of *can* and *could* classified as epistemic.

For deontic uses, the authors give the example, *'I knocked and she said I COULD come in.'*. This is a sequence of tense context, and we would classify the reading of *could* as a grammatical past use of *can* and *'deontic'*. Unlike the authors, we would distinguish between uses of *can* in sequence of tense contexts and past uses such as that in sentence (42) below:

(42) Relatives could visit briefly at prescribed times.

The question is whether the permission, in these examples, was given in the past or in the present at speech time. Thinking of the problem in terms of textual entailment, for example, if someone were trying to figure out if they are allowed to visit the hospital, sentence (42) on its own would not give information about the present possibility, but sentence (43) would.

(43) The administration said that relatives could visit briefly at prescribed times.

For dynamic uses, it is clearest to look at the authors' examples with *can*. These examples include ability, circumstantial, and quantificational uses.

Given that our own results showed higher IAA on examples that were not in SoT contexts, it is likely that adding examples with *can* to the cases with *could* raised rather than lowered the IAA scores, but it is not possible to know since they are reported together.

Using the traditional labels with event-time labels, our raw percentage agreement is higher than that reported in Rupenhoffer & Rehbein (2012), however, our κ is comparable. The difference in our scores has to do with the fact that our sample had a very high number of one single label, specifically, the *'Non-Past B'* label, which led to a lower κ. The higher frequency of non-past Paradigm B readings is the result of focusing on *could*, as uses of *can* only have Paradigm A readings.

Rubinstein et al. (2013) do not provide IAA by modal, so it is difficult to compare their results for a number of modals, many of which present easier classification problems, with our uses of *could*. Also, their ways of collapsing types is so different from ours that there is not basis for comparison.

6.2 Issues in Identifying Sequence of Tense Contexts

The identification of SoT contexts was non-trivial for annotators and required several sessions of training and calibration. Annotators were asked to check whether a sequence of tense context was present. Sequence of tense contexts were most commonly uses of *could* embedded under a past tense verb of saying such as *'said', 'reported', 'told reporters',* and *'announced'*. Annotators took note of whether the use of *could* was in direct quotations or not. If it was in direct quotations, then it was not considered to be a sequence of tense context.

Contexts were determined to be sequence of tense contexts or not, independently of whether or not they were contexts in which the modal auxiliary displayed sequence of tense effects. The reason for this was that annotators unintentionally conflated factors in the analysis when we did not separate the step of determining sequence of tense effects from determining the paradigm reading. Furthermore, sequence of tense contexts lead to a higher degree of ambiguity, especially when only a single sentence in isolation was read. Listing first when a sequence of tense context was present helped annotators consider whether or not they needed to include the grammatical past interpretation as a possible interpretation.

When annotators were uncertain whether or not a sequence of tense context was present, they were instructed to imagine a new sentence with *can* and see if, when in the embedding context of the sentence, it changed to *could*, given the type of embedding verb. At first, annotators found this a bit difficult because, in many dialects of English, uses of *can* that do not have sequence of tense effects are common in sequence of tense contexts, as shown below:

(44) Jess: I can come for dinner.

(45) Pat: Jess said she can come for dinner.

They were asked to substitute *could* to see if it still sounded grammatical when displaying sequence of tense effects, which seemed to help with determination of sequence of tense contexts.

Other examples that posed difficulties were cases in which a past tense verb of saying was followed by a verb of saying in the *'-ing'* form, as in sentence (46) below:

(46) Yeltsin **informed** news sources that he was tired of the plan, **saying** he could not avoid leaving the country. (bold added for emphasis)

as well as examples that were extremely long multi-clausal sentences ending with the phrase, *'sources said'*. Such cases led to occasional errors in the recognition of SoT contexts.

Although past Paradigm B forms occur in sequence of tense contexts, they do not display sequence of tense effects.

(47) Jess: That fool could be the next president.

(48) Pat: Jess said that that fool could be the next president.

It is not the case that sentence (47) is reported using the past Paradigm B form. If it is stated with a past Paradigm B form, it has a different meaning such that the next president has already been selected.

(49) Pat: Jess said that that fool could have been the next president.

Even if the election had already occurred, sentence (49) with the past Paradigm B form does not seem to accurately convey what Jess said at an earlier time in sentence (47).

6.3 Comparison with a Baseline and Sense Reduction

A baseline of the most frequent reading was determined based on the annotators' adjudication of 500 sentences (separate from those on which IAA was calculated). It was determined based on these data that a simple baseline of a *'Non-Past B'* sense, would yield 61.6% accuracy. *'Non-Past B'* is the one label that is uniquely associated with a traditional taxonomic category, therefore, *'Non-Past B'* readings can be referred to as *'possibility'* readings and compared with other studies:

Label	Frequency (out of 500)
non_past_B	61.6%
past_A	21.4%
gram_past	11.2%
past_B	5.2%

TABLE 12 Gold Standard Frequency on a random sample of 500 sentences

The baseline calculation based on the gold standard highlights some important observations. Compared to the preliminary annotation reported in Section 2.1, the frequency of possibility readings increased.

Given the traditional taxonomic classification, possibility reading labels were only given to between 20% and 35% of the data (where results are calculated on $n = 100$ sentences). The difference in the number of possibility readings is partly due to the fact that many of them were present in SoT contexts which, in that round of annotation, were given a separate label, *'report'*.[10] However, the lower reported number is also due to annotator errors, which were significantly reduced in the annotation project using the temporal labels. These finding corroborate the SoT hypothesis which claims that determining a reading of *could* is dependent on recognizing SoT contexts as well as corroborating the Temporally Informed Hypothesis which states that annotators provide better labels when given categories that cause them to pay attention to temporal and aspectual features.

The frequency of non-past Paradigm B examples in the gold standard annotation of 500 sentences is lower than that in the data on which IAA was measured. In these data, they comprised 79% of the data in the cases on which both annotators agreed on the label. This difference could indicate that the percentage of non-past Paradigm B readings in the corpus is lower overall than in the random samples we held out for IAA (where IAA was calculated on $n = 100$ sentences).

During the annotation process, when adjudicating, making annotators aware of the SoT affects often led them to read the sentences differently and reject their initial annotation. That is not to claim that, when reading news reports, human readers are unaware of what a given modal auxiliary means. Rather, it is more likely the case that, they have sufficient world knowledge on the topic to interpret the modal auxiliary in an effortless way.[11] When looking at isolated sentences, however, annotators appear to make sloppy judgments based on lexical content without considering what the temporal and aspectual properties of the sentence convey about the temporal location of an event. It seems that, in the context of a single sentence, the temporal and aspectual properties of the sentences are what carry significant information about modal readings.

It is also worth noting that past Paradigm B forms are very rare in the news corpus, comprising only 5.2% of the 500 sentences. Past Paradigm B forms are easy to recognize due to the presence of the

[10]There may also have been some confusion with the label *'deontic'* as the confusion matrix on the traditional taxonomy annotated along with the temporal labels shows the deontic label nearly disappearing.

[11]Testing whether or not humans tolerate some level of ambiguity or even vagueness in modal auxiliary meaning is an interesting topic for another kind of empirical study.

auxiliary *have* and the perfect form of the verb. They could easily, therefore, be left out of the task.

Dividing the corpus among the remaining three readings suggests an additional cut between non-past Paradigm B readings on one hand and past Paradigm A and grammatically past Paradgim A readings on the other hand. This binary division of the data corresponds to the difference between uses of *could* that support actuality inferences and uses that do not.

6.4 Application to Automated Classification

The primary goal of any classifier is to find meaningful patterns in data. When annotation is being used to train a classifier, the choices made in the annotation process matter. There are many patterns in data sets, only a few of which have applications in current NLP tasks.

The issues with human labeling are only increased for an automated learner, which does not have access to the complex lexical semantic nuances associated with a given sentence that human readers have, nor the degree of world knowledge about the topics in the text.

Using features that are tractable, given state of the art preprocessing tools, makes success more likely.

The considerations discussed so far show that this task has provided two substantial pathways to better solutions to the problem of modal auxiliary classification. The first innovation is that it has essentially reduced a multi-class classification problem to a binary decision separating those instances of sentence with *could* that support actuality inferences from those sentences with *could* that do not.

There are only a small number of modal auxiliary uses that support actuality inferences. They include ability uses of *can* and Paradigm A uses of *could*. The only other modal auxiliary that describes actually occurring events is the use of *would* to describe a past repeated action, as in sentence (50) below:

(50) When I was a kid, we would walk along the riverside scaring the turtles into the water for fun.

There are also actuality inferences in some uses of *would* which describe future events in the past, as in sentences such as (51) below:

(51) I met the person who would one day be my biggest source of support.

None of these readings is common in the news genre, but they do occur in literature and narratives. As far as news corpora are concerned,

the primary source of modal auxiliaries with actuality inferences are uses of *can* and *could*.[12]

The uses of *could* with actuality inferences are most difficult to detect in SoT contexts and such contexts were shown, upon adjudication among annotators, to exhibit genuine ambiguity.

The second innovation is that the difference between the two classes can be formulated almost entirely in terms of grammatical features that are tractable for machine learning. Most of the decision points in the annotators' decision tree correspond to natural language expressions that can be described in terms of their part of speech (pos) tags.[13]

The ambiguities that were not resolved at the sentence level, can, we hypothesize, be resolved by the temporal features of the preceding sentence. In general, detecting grammatical features outside of the sentence window is a more difficult task. However, the tense of the previous sentence would be a helpful feature as possibility or hypothetical readings tend to be in texts about hypothetical situations and ability readings tend to be in texts with past or present tense.[14]

7 Conclusions

Our work demonstrated a novel way to classify uses of *could*. We showed that *could* presents a number of difficulties due to its high number of readings, lack of one single highly dominant sense, and susceptibility to SoT effects.

Sub-groupings of modal auxiliaries that are based on temporal properties are at the heart of the most difficult data in current classifiers. The sub-groupings divide up data in a way that may prove more informative for automated learning. Our temporal semantic categories constitute an instance of reducing labels, but in an intrinsically useful way.

In the work reported in this paper, we demonstrated that grammatical features help to determine the temporal placement of a modal auxiliary and reduce the number of possible readings that a modal auxiliary has.

This work suggests that using a semantically informed divide-and-conquer approach can increase the success of classifying modal auxil-

[12]For more details on the nature of actuality inferences associated with ability *can* and *could*, see the discussion in Moon (forthcoming).

[13]A conscious attempt was made to ensure that the features for machine learning were not dependent on parser output due to the potential introduction of errors, however, given that parsers tend to perform well on newswire, it might be worthwhile to consider using the features on parsed data.

[14]The primary author recorded this pattern of temporal and modal behavior in earlier unpublished work on counterfactual discourses.

iaries according to their readings.

References

Abusch, Dorit. 1997. Sequence of Tense and Temporal De Re. *Linguistics and Philosophy* 20(1):1–50.

Baker, Kathryn, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Use of modality and negation in semantically-informed syntactic machine translation. *Computational Linguistics* 38(2):1–28.

Bhatt, Rajesh. 1999. *Covert modality in non-finite contexts*. University of Pennsylvania.

Bhatt, Rajesh and Roumyana Pancheva. 2006. *Conditionals*, vol. 1 of *The Blackwell Companion to yntax*. Blackwell.

Coates, Jennifer. 1983. *The Semantics of Modal Auxiliaries*. Croom Helm Linguistics Series. London: Croom Helm.

Condoravdi, Cleo. 2002. Temporal interpretation of modals: Modals for the present and for the past. In D. Beaver, S. Kaufmann, B. Clark, and L. Casillas, eds., *The Construction of Meaning*, pages 59–88. CSLI Publications.

Egan, Andy, John Hawthorne, and Brian Weatherson. 2005. Epistemic modals in context. In *Contextualism in Philosophy*, pages 131–168. Oxford University Press.

Hacquard, Valentine. 2006. *Aspects of Modality*. Massachusetts Institute of Technology.

Hacquard, Valentine and Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics* 5:1–29.

Iatridou, Sabine. 2000. The grammatical ingredients of counterfactuality. *Linguistic Inquiry* 31(2):231–270.

Kratzer, Angelika. 1981. The notional category of modality. In H.-J. Eikmeyer and H. Rieser, eds., *Words, Worlds and Contexts*. Berlin: Berlin: de Gruyter.

Kratzer, Angelika. 1991. Modality. In A. von Stechow and D. Wunderlich, eds., *Semantik/Semantics: An International Handbook of Contemporary Research*. Berlin: Berlin: de Gruyter.

Lasersohn, Peter N. 2015. Subjectivity and perspective in truth-theoretic semantics. Draft of 2/11/2015 (cited with permission).

Lassiter, Daniel. 2011. *Measurement and Modality: The Scalar Basis of Modal Semantics*. New York University.

Lewis, David. 1973. *Counterfactuals*. Blackwell Publishers.

MacFarlane, John. 2011. Epistemic modality. In B. Weatherson and A. Egan, eds., *Epistemic Modals are Assessment-Sensitive*. Oxford: Oxford University Press.

Moon, Lori. 2016. *Modal Auxiliary Verbs and Contexts*. Unpublished dissertation, University of Illinois at Urbana-Champaign.

Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English gigaword fourth edition. Web Download: Linguistic Data Consortium.

Romer, Ute. 2004. A corpus-driven approach to modal auxiliaries and their didactics. In J. Sinclair, ed., *How to Use Corpora in Language Teaching*, pages 185–199. John Benjamins.

Rubinstein, Aynat, Hllary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. *Proceedings of IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)* .

Ruppenhofer, Josef and Ines Rehbein. 2012. Yes we can! annotating english modal verbs. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 1538–1545.

Schulz, Katrin. 2007. *Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals*. ILLC Dissertation Series.

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2):143–160.