A Dependency Structure Annotation for Modality

Meagan Vigus, Jens E. L. Van Gysel, and William Croft
Department of Linguistics
University of New Mexico
{mvigus,jelvangysel,wcroft}@unm.edu

Abstract

This paper presents an annotation scheme for modality that employs a dependency structure. Events and sources (here, conceivers) are represented as nodes and epistemic strength relations characterize the edges. The epistemic strength values are largely based on Saurí and Pustejovsky's (2009) FactBank, while the dependency structure mirrors Zhang and Xue's (2018b) approach to temporal relations. Six documents containing 377 events have been annotated by two expert annotators with high levels of agreement.

1 Introduction

Representing modality is fundamental to creating a complete representation of the meaning of a text. Modality characterizes the reality status of events, i.e. whether they occur in the real world, or in any number of non-real 'worlds'.

In this paper, we develop an annotation scheme that builds on Saurí and Pustejovsky's (2009) FactBank annotation scheme and Zhang and Xue's (2018b) temporal dependency structures. Although we have only applied this annotation to texts in English, we intend for it to be applicable cross-linguistically (see Van Gysel et al. 2019).

Like FactBank, we combine modality and polarity values and relate both back to a source (or, in our terms, conceiver); the modality/polarity values represent the source's perspective on an event. We propose two main innovations to FactBank's annotation scheme: the interpretation of epistemic strength values in the domains of deontic and dynamic modality, and the representation of modality in a dependency structure.

Modality is generally taken to encompass epistemic, deontic, and dynamic modality (e.g., Palmer 2001). Epistemic modality corresponds most straightforwardly to factuality in that it characterizes whether an event occurs in the real world.

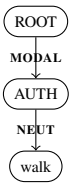


Figure 1: *Mary **might*** HAVE WALKED *the dog.*

We propose that epistemic modality may be interpreted in the domain of deontic modality as degree of predictability (see 3.2.2) and within the domain of dynamic modality as the strength of a generalization over instances (see 3.2.3).

The second main innovation of this paper is the representation of modal annotation as a dependency structure. The dependency structure is a directed, acyclic graph with conceivers and events as nodes and edges between the nodes labelled with epistemic strength values. A simple example of this can be seen in Figure 1; Figure 1 shows that the author has neutral epistemic stance towards the occurrence of the walking event.

This modal dependency structure is based largely on Zhang and Xue's (2018b) temporal dependency tree structure. Structuring the annotation of temporal relations as a dependency tree allows for the same values to be used for temporal relations between events, between time expressions, and between an event and a time expression. This leads to a perspicuous representation of the temporal structure of an entire document.

For modality, the dependency graph structure allows for the nesting of modal values that is necessary to represent certain types of linguistic constructions (see 3.3). The dependency structure also allows for the explicit representation of scope relations between modality and negation. Most of the time, the dependency graph for modality is also a tree: each node only has one parent. However, there are rare cases that require a single event to

Proceedings of the First International Workshop on Designing Meaning Representations, pages 182–198
Florence, Italy, August 1st, 2019

have two parents in the graph; see 3.3.

A dependency structure for modal annotation has another advantage: it closely mirrors the mental spaces theory of modality (Fauconnier, 1994, 1997). This allows for the insights of the mental spaces theory of modality to be straightforwardly imported into our modal dependency structure (see 2.2).

The modal dependency annotation scheme was tested on six documents[1] containing 108 sentences. A total of 377 events were annotated for modality by two expert annotators. Agreement scores were relatively high and similar to those reported in Zhang and Xue (2018b).

2 Background

2.1 Related work

Modality, factuality, certainty, or veridicality of statements in text has been addressed in a variety of ways in the computational linguistics literature (see Morante and Sporleder 2012). In this section, we briefly survey some of the annotation schemes intended to capture modality and polarity distinctions in general-domain texts (see also Nissim et al. 2013; Lavid et al. 2016; Prasad et al. 2008). Although we focus on manual annotation, there have also been automatic annotations of modal information (e.g., Saurí and Pustejovsky 2012, Baker et al. 2010).

Wiebe et al. (2005) focus on the annotation of opinions, emotions, and sentiments, in addition to modality. Importantly, Wiebe et al. (2005) introduce the notion of *nested sources*, including the representation of all in-text sources as nested underneath the author. This notion has been widely adopted and we adopt it in the modal dependency structure.

Rubin et al. (2005) and Rubin (2007) annotate certainty in a corpus of news articles. They annotate four dimensions: level of certainty, perspective (i.e., source), focus (abstract vs. factual), and time reference. Level of certainty is divided into a four-way distinction (absolute, high, medium, and low), however Rubin (2007) reports low inter-annotator agreement for this four-way distinction and suggests a binary distinction may lead to higher agreement. De Marneffe et al. (2012), however, find that annotators actually reached higher agreement scores using FactBank's three-way modality distinctions (see below) as opposed to using a smaller number of distinctions.

Matsuyoshi et al. (2010) annotate seven modal categories: source, time, conditional, primary modality, actuality, evaluation, and focus. Conditional distinguishes between propositions with conditions and those without. Primary modality distinguishes between a number of fine-grained modality categories (e.g., volition, wish, imperative). Their actuality category refers to level of certainty; evaluation refers to an entity's attitude towards an event.

Ruppenhofer and Rehbein (2012) annotate the MPQA corpus (Wiebe et al., 2005) with modality information, focusing on sense disambiguation of grammaticalized modal verbs. In addition, their annotation scheme identifies the modalized Proposition, the Source, and the Link that introduces the source. They focus on distinguishing the modality type (epistemic, deontic, etc.) as opposed to the degree of likelihood, the focus of the current paper. Ruppenhofer and Rehbein (2012) are more restricted than the current scheme in that they limit their annotations to grammaticalized modal verbs.

Rubinstein et al. (2013) report on a language-independent modal annotation that has been applied to the MPQA corpus (Wiebe et al., 2005). Rubinstein et al. (2013) identify and annotate "modal expressions" for modality type, polarity, propositional arguments, source, and a few other categories. They find that annotators are only able to reliably distinguish between rather coarse-grained modality types, essentially epistemic vs. root modality (what they call non-priority vs. priority). Similar to Ruppenhofer and Rehbein (2012), Rubinstein et al. (2013) focus on the type of modality, but do not annotate the propositional arguments with their degree of likelihood (the focus of the current scheme).

FactBank (Saurí and Pustejovsky, 2009) presents a corpus annotated with information about event factuality. They distinguish three levels of factuality: certain (CT), probable (PR), and possible (PS). These interact with a binary polarity distinction, positive (+) and negative (-).

FactBank also introduces an unspecified value (U) for both factuality and polarity. FactBank uses the unspecified values for cases where the factual status of an event is not clear. This can be because the source does not know the factual sta-

[1] These documents were excerpted from Strassel and Tracey (2016), Garland et al. (2012), and *The Little Prince* (de Saint-Exupéry and Woods, 1943).

tus of an event (e.g., *John does not know whether Mary came.*; Saurí and Pustejovsky 2009, 247; compare the '?' mental space relation in Fauconnier 1994, 86) or because the source does not communicate the polarity of an event (e.g., *John knows whether Mary came.*; Saurí and Pustejovsky 2009, 247; compare the '!' mental space relation in Fauconnier 1994, 86). In total, FactBank distinguishes eight factuality values: CT+, CT-, PR+, PR-, PS+, PS-, CTU, and UU.

As mentioned above, FactBank represents these values as tied to a particular perspective, or source. When a source is not explicitly mentioned in the text, the author of the text is the implied source. FactBank also allows for the nesting of sources (as in Wiebe et al. 2005); whenever a source is mentioned in the text, it is annotated as nested underneath the author.

De Marneffe et al. (2012) annotate pragmatic factuality information on top of the more lexically-based factuality information from FactBank. Similarly, this paper proposes an annotation scheme for modality based on the full context of sentences, and not the general meaning of lexical items.

2.2 Mental spaces

Mental space theory was developed by Fauconnier to solve problems of referential opacity and presupposition "projection" (Fauconnier 1994, 1997; see also McCawley 1993). These problems arise because referents and presupposed events may exist only in a non-real *mental space*. A mental space is a representation of alternative realities to the real world—more precisely, the world of the author's beliefs. Mental spaces present alternative realities as cognitive, that is, in the mind of a conceiver, rather than as metaphysical entities, as is done in possible worlds semantics. Mental spaces have entities that are counterparts to real entities (though some may not have real world counterparts), with associated properties and events that are different from those of the real world entities.

The alternative realities represented by mental spaces include both events whose factuality is less than certain, including negative events, which are typically expressed by grammatical modality and negation; and events that are believed, desired, feared, and so on by a conceiver, which are typically expressed by propositional attitude, desiderative, and other such predicates. These alternative realities give rise to the paradoxes in reference and presupposition that interested Fauconnier. We are, however, interested in using the mental space representation to model modality, negation, and predicates that give rise to alternative realities. All such constructions are *space builders* in Fauconnier's terms.

Mental spaces can be nested within other mental spaces. For example, the space representing a person's desire to go to Florence is nested in the space representing that person's beliefs. The nested mental space structure allows one to capture scope relations between modality, propositional attitude predicates, and negation. In fact, the dependency graph structure of nested mental spaces is a more powerful representation than linear scope relations and is able to handle the sorts of semantic and pragmatic problems that Fauconnier analyzes in his work. The dependency structure of mental space relations allows us to adapt the temporal dependency annotation scheme of Zhang and Xue (2018b) to the annotation of modality and related concepts.

3 Modal dependency structure

The modal dependency structure consists of three parts: conceivers/sources, events, and the relations between them. Section 3.1 describes the types of nodes in the dependency structure and 3.2 describes the types of edges.

3.1 Nodes in the modal dependency structure

There are two distinct types of nodes in the modal dependency structure: conceivers and events. Events may have either conceivers or events as parents; conceivers only ever have other conceivers (or, ROOT) as parents. That is, conceivers are never the children of events.

3.1.1 Conceivers

The mental-level entities whose perspective on events is modeled in the text are called CONCEIVERS. Each text will automatically have at least one AUTHOR conceiver node, representing the perspective of the creator of the text. Texts with multiple creators (e.g., dialogues) will have multiple AUTHOR nodes.

When the author models the mental content of other entities, those entities are also represented as conceiver nodes in the dependency structure. Certain types of predicates inherently involve conceivers: report, knowledge, belief, opinion, doubt, perception, and inference (Saurí and Pustejovsky,

2009, 236). For example, in *Mary thinks the cat is hungry*, the author is asserting something about the content of Mary's attitudes and beliefs. Therefore, MARY is identified as a conceiver and added as a node in the graph.

In contrast to FactBank, we introduce conceiver nodes for deontic events (e.g., volition, intention). FactBank excludes them because they express an attitude that is not "epistemic in nature" (Saurí and Pustejovsky, 2009, 237). However, we take a broader view of sources as conceivers whose mental content is expressed in the text; a person's desires or intentions are based on their own set of beliefs, and not the author's beliefs (McCawley 1993, 421; Fauconnier 1994, 93). For deontic events, this allows us to annotate the strength of likelihood that the future event will occur based on the conceiver's mental attitude. Wiebe et al. (2005) also annotate sources for deontic events.

Also following Wiebe et al. (2005), we represent conceiver nodes as children of the AUTHOR node. Another conceiver's mental space is always mediated by the author's perspective. For example, in *Mary thinks the cat is hungry*, the author is attributing a belief to Mary; as readers, we don't have direct access to Mary's mental content, only to the author's perspective on Mary's beliefs. Therefore, the MARY node is represented as a child of the AUTHOR node.

There may be an indefinite number of nested conceivers. For example, *Mary said that Henry told her that John thinks the cat is hungry* has four conceivers (including the author). The JOHN conceiver node is nested underneath the HENRY node, which is in turn nested underneath the MARY node; finally, the MARY node is a child of the AUTHOR node.

Although conceivers are prototypically mental-level entities, conceiver nodes can also be used to represent the "world" in which a particular event takes place in the case of stories, drawings, movies, etc. For example, in *Aeneas flees Troy in The Aeneid*, AENEID is identified as a conceiver; all events in the story, such as *flee*, are nested underneath AENEID.

3.1.2 Events

The other type of node in the modal dependency structure represents the events themselves. We largely follow TimeML's event identification criteria (Pustejovsky et al., 2005).

The only semantic type of event which we exclude from our modal dependency structure are events that attribute beliefs to a conceiver (e.g., *think, believe*). These events correspond straightforwardly to the edges in the modal dependency structure (see 3.2), and therefore they are not represented as nodes. For the same reason, we also do not represent grammaticalized modal auxiliaries (e.g., *may, must*) as nodes in the dependency structure.

3.2 Edges in the modal dependency structure

As mentioned in 1, the edges in the modal dependency structure correspond to combined epistemic strength and polarity values. These characterize the type of mental space in which a particular event holds. Edges can link two events, two conceivers, or a conceiver and an event.

In a cross-linguistic study drawing on data from fifty languages, Boye (2012) finds that three levels of epistemic support are sufficient to characterize epistemic modal systems across languages. That is, languages tend to have forms that distinguish three levels of epistemic support. Boye (2012) uses the term "support" to refer both to epistemic modality proper and the combination of evidential and epistemic modality (see 3.2.1). Following Boye (2012), we label our values FULL, PARTIAL, and NEUTRAL. Since we extend our values outside of prototypical epistemic and evidential modality, we refer to these values as characterizing epistemic "strength". These three values correspond to FactBank's CERTAIN, PROBABLE, and POSSIBLE values.

Also like FactBank, we combine these values with a binary polarity distinction (POSITIVE/NEGATIVE) for a total of six values. These strength/polarity values represent the modality as scoping over negation. For less grammaticalized forms that express combinations of modality and negation, the dependency structure represents the scope relations between the two.

The combined modality-polarity values are shown in Table 1. These values characterize the likelihood that a particular event occurs (or does not occur) in the real world. The lexical item in the examples that expresses the epistemic strength of the sentence is in bold. For the POS value, the simple declarative sentence in English conveys full positive epistemic strength; this is very common cross-linguistically (Boye, 2012).

Epistemic strength is generally only used to de-

Label	Value	FactBank	Definition	Example
POS	full positive	CT+	complete certainty that event occurs	*The dog* BARKED.
PRT	partial positive	PR+	strong certainty that event occurs	*The dog* ***probably*** BARKED.
NEUT	positive neutral	PS+	neutral certainty that event does/n't occur; expressed positively	*The dog* ***might*** *have* BARKED.
NEUTNEG	negative neutral	PS-	neutral certainty that event does/n't occur; negation expressed	*The dog* ***might not*** *have* BARKED.
PRTNEG	partial negative	PR-	strong certainty that event does not occur	*The dog* ***probably didn't*** BARK.
NEG	full negative	CT-	complete certainty that event doesn't occur	*The dog* ***didn't*** BARK.

Table 1: Strength values

scribe phenomena like those in Table 1: the factuality of a single instance of a specific event in the past or present. We use the notion of epistemic strength to characterize evidential justification, deontic modality, and dynamic modality.

Although epistemic strength is interpreted slightly differently in these domains, it still refers to the likelihood of occurrence of the event in question in the real world. It is important to note that the modal dependency structure itself does not distinguish between episodic, deontic, or dynamic events. However, the modal annotation scheme may be used in conjunction with other annotations which do distinguish between these types of events (e.g., temporal or aspectual annotation).

3.2.1 Evidential justification

Following Boye (2012) and Saurí and Pustejovsky (2009), we characterize evidential justification in terms of epistemic support.

Boye (2012) finds that there is cross-linguistic evidence for lumping epistemic support and evidential justification together into the same relations. Specifically, languages may encode direct evidential justification (sensory perception) with the same forms as full epistemic support; indirect justification (hearsay, inferential) may be encoded by the same forms as partial epistemic support.

Example 1 shows how direct and indirect justification correspond to epistemic support.

(1) a. *I* ***saw*** *Mary* FEED *the cat.*

b. *Mary* ***must*** *have* FED *the cat.*

In 1a, the author has direct knowledge of the feeding event, by way of witnessing it. Therefore, *feed* would be annotated with POS strength. In 1b, however, *must* signals that the author is inferring that the feeding event occurred without direct, perceptual knowledge. Therefore, *fed* in 1b would be annotated with PRT strength.

3.2.2 Deontic modality

We analyze deontic modality (e.g., desires, intentions, demands) as a subtype of future events, since the event that is desired, demanded etc. will take place in the future if it takes place at all. We group together deontic events and simple assertion of future events as 'future-oriented' events.

In the modal dependency structure, we interpret epistemic strength within the future-oriented domain as degree of predictability, rather than degree of factuality, because future events are unverifiable at the present moment.

Example 2 shows the three degrees of epistemic strength within the future-oriented domain.

(2) a. *Bill* ***will*** DRIVE *to Pisa.*

b. *Bill* ***is planning*** TO DRIVE *to Pisa.*

c. *Bill* ***wants*** TO DRIVE *to Pisa.*

Example 2a, the plain future, represents the highest degree of predictability for future-oriented events; therefore, this corresponds to FULL strength. Intention, as in 2b, is annotated as PARTIAL strength in the future-oriented domain: once an agent forms an intention, the event is likely to occur. Desire, as in 2c, corresponds to NEUTRAL strength: one may or may not act on one's desires.

Future-oriented events can also occur in the past (i.e., the future-in-the-past), as in example 3.

(3) a. *Bill* ***would*** DRIVE *to Pisa (the next morning).*

b. *Bill* ***was planning*** TO DRIVE *to Pisa.*

c. *Bill* ***wanted*** TO DRIVE *to Pisa.*

Akin to 2, the future-in-the-past can also occur with different strengths. That is, 3a implies that the driving event happened, i.e. FULL strength.[2] Example 3b expresses past intention, which opens

[2]The main clause use of *would* is not the same as *would* occurring in conditional constructions (Fillmore, 1990).

up the possibility that the driving event didn't actually happen; this corresponds to PARTIAL strength. In 3c, only a past desire is expressed, without any indication whether or not the driving event actually took place; this is NEUTRAL strength.

3.2.3 Dynamic modality

Epistemic strength is generally not considered to apply to dynamic modality or generic statements because they do not refer to a specific instance of an event, but a generalization over instances.

In this paper, we tentatively propose that dynamic modality and generics can be subsumed under the same analysis as generalizations that can be mapped onto actual, episodic events.[3] The two levels of dynamic modality (possibility and necessity) combined with generics creates a three-way distinction that can be characterized in terms of strength. Dynamic possibility, as in *Owls can hunt at night*, corresponds to epistemic possibility, i.e. NEUTRAL strength. Dynamic necessity, as in *Owls must hunt at night*, corresponds to epistemic necessity, i.e. FULL strength. Generic events, as in *Owls hunt at night*, represent a generalization between "possibly" and "necessarily"; generics express that something occurs "usually" or "normally". Therefore, we analyze generics as PARTIAL strength.

The correspondence between strength values with episodic and generic events can also be thought of in these terms: a FULL strength generic can be falsified by one negative episodic event, a NEUTRAL strength generic is verified by one positive episodic event, and a PARTIAL strength generic cannot be falsified by one negative episodic event, but there must be enough relevant episodic events to infer that the event is typical or characteristic.

3.2.4 Edges between conceiver nodes

Edges between conceiver nodes are characterized by the same set of strength distinctions. That is, just as conceivers may express different strengths towards events, they also may express different strengths towards the modeling of another conceiver's mental content. This can be seen in Table 2. The epistemic strength values correspond to the relation between the AUTHOR conceiver node and the MARY conceiver node.

Value	Example
POS	*Mary knows the cat ate.*
PRT	*Mary **probably** knows the cat ate.*
NEUT	*Mary **might** know the cat ate.*
NEUTNEG	*Mary **might not** know the cat ate.*
PRTNEG	*Mary **probably doesn't** know the cat ate.*
NEG	*Mary **doesn't** know the cat ate.*

Table 2: Edges between conceiver nodes

3.2.5 Summary of edge values

Extending epistemic strength to cover evidential justification, future likelihood, and strength of generalization over instances allows us to use a single set of distinctions to characterize (and annotate) events in different modal domains.

3.3 Dependency structure

The second main innovation in this annotation scheme is the representation as a dependency structure, as opposed to assigning a single modal value to an event. The dependency structure allows us to nest modal strengths between events. This can be seen in example 4.

(4) *Mary **might need** TO CHECK the weather.*

This example contains two modal expressions: epistemic *might* and deontic *need*. That is, *might* expresses a NEUTRAL epistemic stance towards the needing event; *need* expresses a PARTIAL epistemic stance towards the checking event.

If we were to assign a single annotation value to *check*, it is not clear if this should be NEUT from *must* or PRT from *need*. The dependency structure allows us to explicitly represent this nesting of strength values. This can be seen in Figure 2. Here, *check* is represented as the child of *need*, with a PRT relation. The *need* event is represented with a NEUT relation to the AUTHOR node.

Example 5 illustrates another case where representing the nesting of modal relations between events is necessary.

(5) *I'll **probably allow** EATING in the classroom this year.*

Here, *probably* indicates PRT strength, whereas *allow* indicates NEUT strength; see Figure 3.

As mentioned in 1, there are rare cases where a single node has two parents in the dependency

[3] We also tentatively propose that all generics are represented with their own node in the dependency graph. That is, *Owls hunt at night* would require two nodes: one for the generic and one for *hunt*. This is necessary in order to capture situations in which epistemic modals scope over the generic, e.g. *Owls might hunt at night*. This includes dialects of English in which double modals (e.g., *might can*) occur.

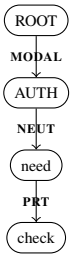


Figure 2: Strength nesting: *Mary might need to check the weather.*

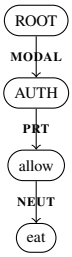


Figure 3: Strength nesting: *I'll probably allow eating in the classroom this year.*

graph. The clearest example of this is with *know*, as in 6 below.

(6) *Mary* ***knows*** *the cat* ATE *breakfast.*

The issue here is that *know* tells us something both about Mary's beliefs and the author's beliefs. That is, *know* in 6 implies that the author shares Mary's beliefs about the eating event. Thus, the eating event is represented as a child of both the AUTH node and MARY node; see Figure 4.

4 Annotation

4.1 Annotation procedure

The modal dependency structure annotation proceeds in three passes. Disagreements were resolved between each pass. In the first pass, the events that will be annotated for modality are identified. This is done based largely on TimeML's (Pustejovsky et al., 2005) event identification;

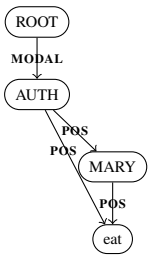


Figure 4: Multiple parents: *Mary knows the cat ate breakfast.*

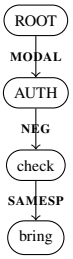


Figure 5: Same space: *Mary didn't check the weather or bring a map.*

events are identified based on semantic criteria and not morphosyntactic structure or part of speech.

The next pass involves setting up the modal 'superstructure'. This is akin to the identification of time expressions in Zhang and Xue (2018b); it builds the top of the graph, which applies to an entire document. At the top of each graph is a ROOT node. For modality, there is also always an AUTHOR conceiver node. Underneath the AUTHOR conceiver node are nodes for all of the other conceivers in the text. As mentioned in 3, the edges between conceiver nodes are distinguished by the epistemic strength relations.

The third pass involves the modal annotation. For each event identified in the first pass, annotators select a parent node (either another event or a conceiver) and the appropriate strength relation between the parent and child nodes.

In addition to the strength relations in Table 1, we introduced a Same Space (SAMESP) relation between nodes. The SAMESP annotation indicates that two events hold in the same mental space, i.e. they have the same strength relation from the same conceiver node. For example, in *Mary didn't check the weather or bring a map*, both *check* and *bring* have a NEG relation to the AUTHOR node. This would be annotated with a NEG relation between *check* and MARY and a SAMESP relation between *bring* and *check*; see Figure 5.

4.2 Current Implementation

The modal dependency structure annotation has been tested on six documents, containing 108 sentences with 377 identified events. These documents have been annotated by two expert annotators. Please refer to the supplementary material for annotated sections from these documents, including their representation as a dependency graph. In addition to the manually-created dependency graphs, the supplementary material also contains graphs generated automatically with the Abstract

Pass	Measure	News	Narr.	Forum	Total
Event ID	Precision	0.95	0.95	0.98	0.94
	Recall	0.92	0.92	0.87	0.93
	F-score	0.94	0.93	0.92	0.93
Conceiver	Precision	0.9	0.86	1	0.91
	Recall	0.82	0.75	0.64	0.77
	F-score	0.86	0.80	0.78	0.83
Event space	Precision	0.93	0.84	0.78	0.88
	Recall	0.93	0.83	0.78	0.88
	F-score	0.93	0.83	0.78	0.88

Table 3: IAA for modal annotations

Meaning Representation Reader (Pan et al., 2015).

The inter-annotator agreement scores for each of the three annotation passes are shown in Table 3. These agreement scores reflect only true disagreements between annotators; they disregard cases in which annotators used a different annotation to represent the same modal analysis.[4]

The annotated documents represent three different genres: news stories, narratives, and discussion forums. The first row shows precision, recall, and F-score for the first pass, event identification, in all three genres, following Zhang and Xue (2018a). The middle row shows the same measures for the second pass, the identification of the conceiver nodes in the superstructure; the bottom row shows these measures for the third pass, the mental space annotation of each event - 228 in the news genre, 85 in the narrative genre, and 64 in discussion forum texts.

Zhang and Xue (2018b) report the following F-scores (for news and narrative respectively): .94, .93 for event recognition, .97, 1 for timex recognition, and .79, .72 for event relations.

Our event identification F-scores are identical to Zhang and Xue (2018b) in the news and narrative genres. Their timex recognition corresponds to our modal superstructure (essentially conceiver recognition). Our superstructure F-scores are noticeably lower than their timex recognition F-scores. We believe this is because of the relative difficulty of identifying when an entity's mental content is modeled vs. when a linguistic expression refers to a locatable point in time. See 4.3 for a more detailed discussion.

Importantly, our event annotation F-scores are largely consistent with, if not slightly higher than Zhang and Xue (2018b) report for their event relation scores. This suggests that annotators are able to consistently assess the epistemic strength relations and relevant conceivers in a text and uniformly model them in a dependency structure.

[4]The SAMESP label led to cases where the annotators had the same strength relation underneath the same conceiver (i.e., the same modal analysis), but one annotator notated it with SAMESP. These types of notational errors made up 34% of total errors. Therefore, we have removed the SAMESP label from the modal annotation scheme.

4.3 Modal error analysis

This section will discuss and exemplify the types of disagreements that arose between annotators for the second and third passes.

Error type	Percentage of total
Lexical item	53%
Childless conceiver	29%
Different parent	12%
Co-referential nodes	6%

Table 4: Conceiver errors

Table 4 shows the types of errors that arose in the second pass. The most common disagreement between annotators was whether a particular lexical item required the introduction of a conceiver node in the superstructure. That is, annotators disagreed about whether a particular lexical item represented the author's modeling of another entity's mental content, as in 7.

(7) *Christie is being set up on this one and* THE LEGISLATOR ***called his bluff***.

The issue here is whether the idiom *call...bluff* invokes the mental content of its subject, here *the legislator*. That is, is the author simply reporting an event, or is the author ascribing mental content (e.g., the knowledge that Christie is bluffing) to *the legislator*? Like many of the disagreements based on which lexical items invoke conceivers, this seems like a case of genuine ambiguity.

The second most common type of superstructure error was whether childless conceivers were represented in the modal superstructure. Annotators differed on whether they added nodes to the superstructure for conceivers whose nodes would not have any events as children; this is shown in 8.

(8) PEOPLE ***seeking bargains*** *lined up at a state-run food store in La Paz on Thursday...*

Here, it is clear that *seek* requires modeling the mental content of another entity, *people*. However, there would be no event represented as a child of the PEOPLE conceiver node, since the object of *seek* is not an event. For subsequent annotation,

we have decided that conceivers should be represented in the modal superstructure, even if they won't have any events as children; this should alleviate these types of disagreements.

The different parent disagreements refer to cases where annotators identified the same entities as conceivers, but differed on whether they were children of the AUTHOR or another conceiver in the text. Finally, there was disagreement between annotators based on whether entities mentioned in the text were co-referential or not. That is, annotators agreed about when conceiver nodes were necessary, but disagreed about whether two conceiver instances referred to the same entity.

For the third pass, the modal annotation of events, Table 5 shows the types of disagreements between annotators.

Error Type	Percentage of total
Lexical item	34%
Space scope	23%
Conceiver scope	16%
Space type	14%
Miscellaneous	9%
Annotator error	4%

Table 5: Event annotation errors

The most common disagreements concern the strength of particular lexical items, as in 9.

(9) *Lerias* ***called for*** *more rescuers* TO COME *to the site...*

The issue here is the strength that *call for* implies for *to come*; annotators disagreed on whether *to come* has PRT strength or NEUT strength. The frequency of this type of disagreement can probably be diminished by training annotators with more specific guidelines for each strength relation; however, some of these types of disagreements will likely be inevitable.

Space scope disagreements refer to cases where annotators disagreed about whether a particular event belongs in the same mental space as the preceding event in the text. This is shown in 10.

(10) *In the book it said: "Boa constrictors swallow their prey whole, without chewing it. After that they are not able* TO MOVE *..."*

Both annotators agreed that *swallow* and *chewing* belong in a "usually", i.e. PRT strength, generic space. Annotators also agreed that *not able* indicated NEG strength of the *to move* event. The disagreement is whether the PRT strength generic scopes over the *to move* event. That is, is *to move* the (direct) child of BOOK or the child of an event in the PRT generic space? Some cases like these may be resolved by more detailed guidelines on determining the scope of spaces over events.

Similarly, the scope of conceivers over events was a source of disagreement. This generally occurred with indirect speech predicates, as in 11.

(11) *Lerias called for more rescuers to come to the site to help look for bodies as heavy earth moving equipment could not* WORK *in the mud...*

Here, annotators disagreed on whether LERIAS or AUTHOR was the source for the *work* event. These errors appear to represent textual ambiguity.

Space type errors refer to cases where annotators disagreed on whether an event was in an episodic, generic, or future-oriented space. Although the modal annotation scheme does not directly distinguish these different space types, annotators' interpretation was evident in the strength relation chosen, as in example 12.

(12) *Military helicopters were able* TO REACH *the area despite heavy clouds...*

Annotators disagreed about whether this sentence represents a NEUT strength "possibility" generic, based on the use of *able*, or whether *to reach* represents full POS strength because the past tense implies that the event did occur.

5 Conclusion

A modal annotation scheme structured as a dependency graph, like the temporal annotation scheme of Zhang and Xue (2018b), captures the complexity of modal relations (mental space structure) in sentences and documents with a relatively simple annotation: each event has a single annotation of the event's modal relation to another event or a conceiver, not unlike the single annotation of an event's temporal relation to another event or a time expression. The pilot annotation indicates that this annotation scheme is relatively easy to implement, at least as much as the annotation of temporal dependency structure.

Acknowledgments

This research was supported in part by grant 1764091 by the National Science Foundation to the last author.

References

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC '10)*, pages 1402–1407, Valette, Malta. European Language Resources Association (ELRA).

Kasper Boye. 2012. *Epistemic meaning: A crosslinguistic and functional-cognitive study*, volume 43 of *Empirical Approaches to Language Typology*. De Gruyter Mouton, Berlin.

Gilles Fauconnier. 1994. *Mental spaces*, 2 edition. Cambridge University Press, Cambridge.

Gilles Fauconnier. 1997. *Mappings in thought and language*. Cambridge University Press, Cambridge.

Charles F. Fillmore. 1990. Epistemic stance and grammatical form in English conditional sentences. In *Papers from the 26th Regional Meeting of the Chicago Linguistic Society*, pages 137–62, Chicago. Chicago Linguistic Society.

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, and Haejoong Lee. 2012. Linguistic resources for genre-independent language technologies: user-generated content in bolt. In *Workshop Programme*, page 34.

Julia Lavid, Marta Carretero, and Juan Rafael Zamorano-Mansilla. 2016. Contrastive annotation of epistemicity in the multinot project: preliminary steps. In *Proceedings of the ISA-12, Twelfth Joint ACL-ISO Workshop on Interoperable Semantic Annotation, held in conjunction with Language Resources and Evaluation Conference*, pages 81–88.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38:301–333.

Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating event mentions in text with modality, focus, and source information. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC '10)*, pages 1456–1463, Valette, Malta. European Language Resources Association (ELRA).

James D. McCawley. 1993. *Everything that Linguists have Always Wanted to Know about Logic* (*but were too afraid to ask)*. University of Chicago Press, Chicago.

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

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

F. R. Palmer. 2001. *Mood and Modality*. Cambridge University Press, Cambridge.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1130–1139, Denver, Colorado. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39:123–164.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *NAACL 07: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Morristown, NJ, USA. Association for Computational Linguistics.

Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2005. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *Information Retrieval Series*, pages 61–76. Springer-Verlag, New York.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 38–46, Potsdam, Germany. Association for Computational Linguistics.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? annotating the senses of English modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*.

Antoine de Saint-Exupéry and Katherine Woods. 1943. *The Little Prince*. Harcourt, Brace & World, New York.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38:261–299.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Jens E. L. Van Gysel, Meagan Vigus, Pavlina Kalm, Sook–kyung Lee, Michael Regan, and William Croft. 2019. Cross-lingual semantic annotation: Reconciling the language-specific and the universal. In *First Workshop on Designing Meaning Representations, Association for Computational Linguistics*. Association for Computational Linguistics.

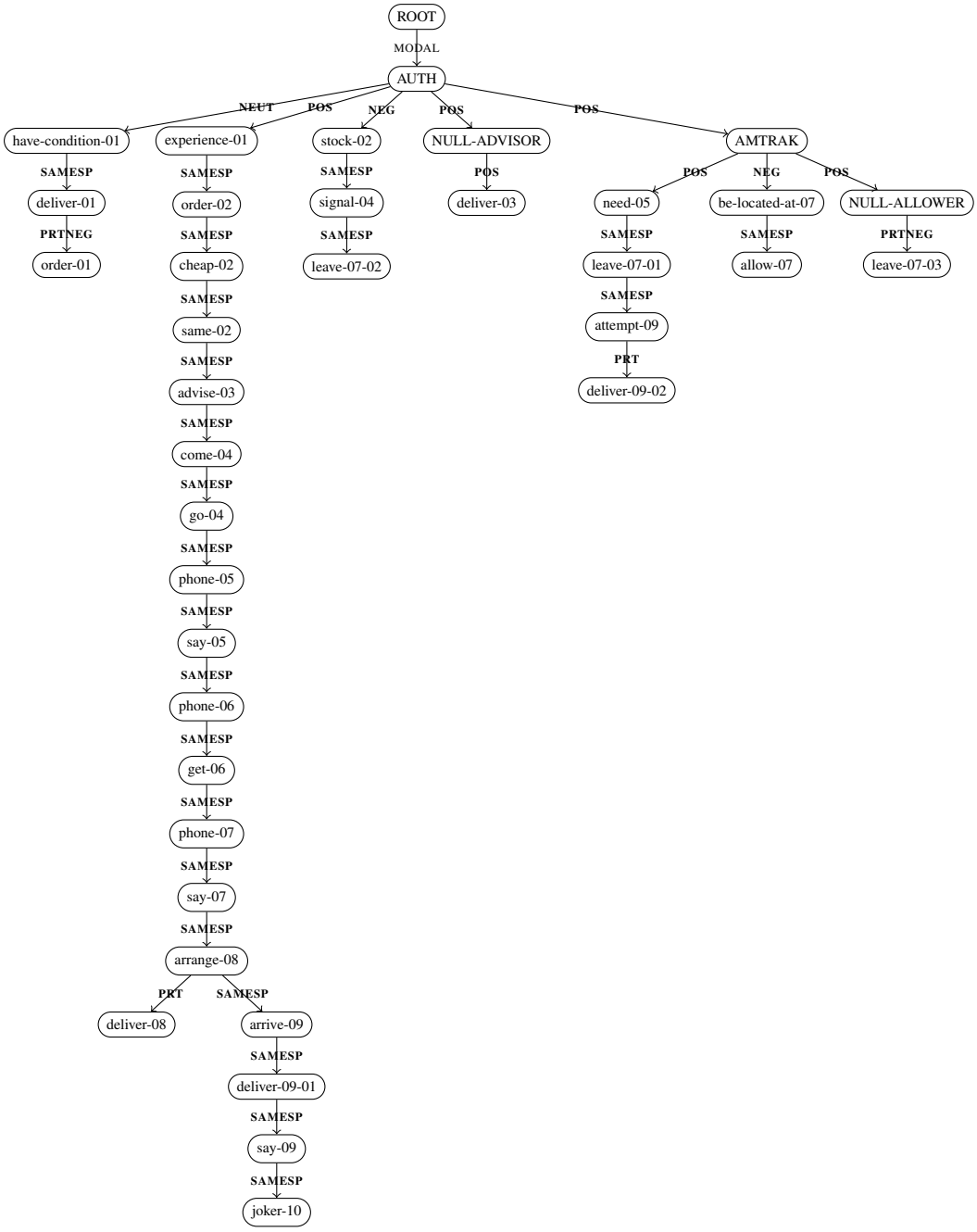
Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Yuchen Zhang and Nianwen Xue. 2018a. Neural ranking models for temporal dependency structure parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3339–3349.

Yuchen Zhang and Nianwen Xue. 2018b. Structured interpretation of temporal relations. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

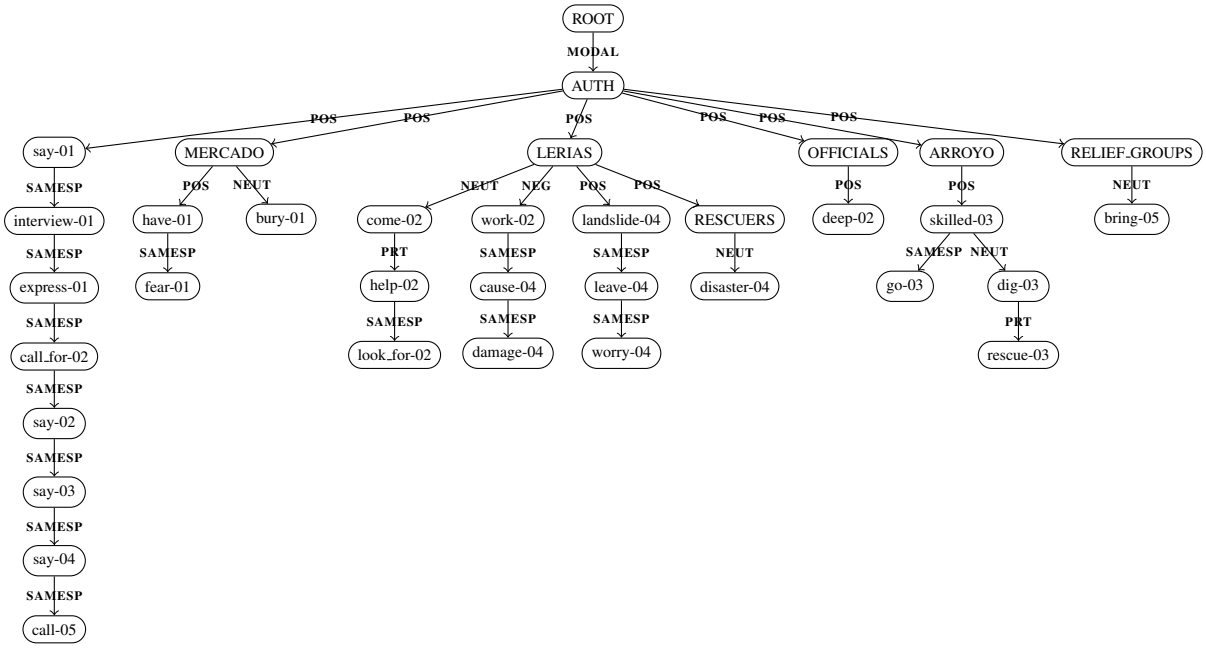
Supplementary Material

Genre: Discussion Forum



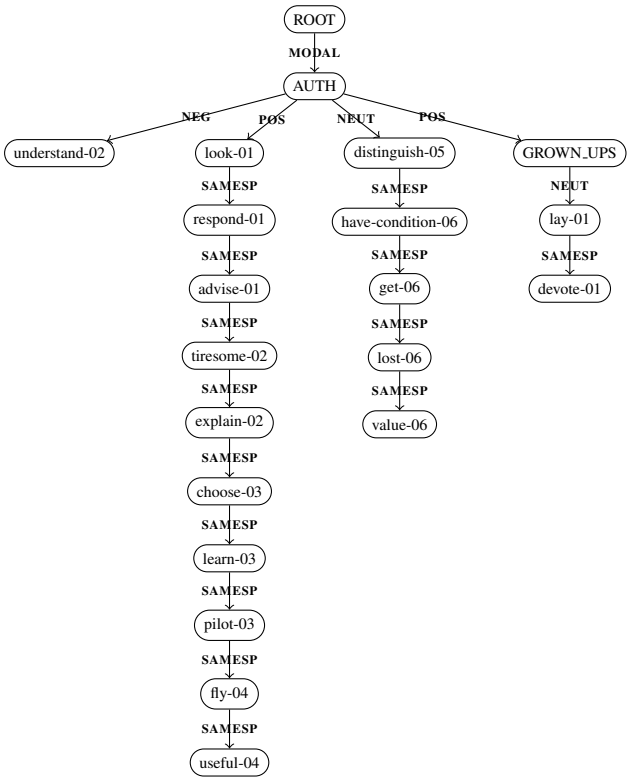
1 Don't **order** anything online if Amtrak **are delivering** it - here's my **experience**.
2 **Ordered** a 32" TV online, **cheaper** than Argos-who didn't **have** it **in stock**-but with the delivery charge the cost **was the same**.
3 **Advised** that it **would be delivered** by Amtrak on Tuesday.
4 Tuesday **came** and **went**, no **sign**.
5 **Phoned** Amtrak on Wednesday, "we **need** a consignment number".
6 **Phoned** online company and **got** it.
7 **Phoned** Amtrak "a card **was left** on Tuesday as you **weren't there**" (no it **wasn't** of course), and "we're not **allowed to leave** it with a neighbour".
8 **Arranged** for another **delivery** on Saturday.
9 **Arrived** home yesterday-it **had been delivered** next door yesterday, with a card **saying** this was their first **attempt** at **delivery**...
10 What **a bunch of jokers**.

Genre: News



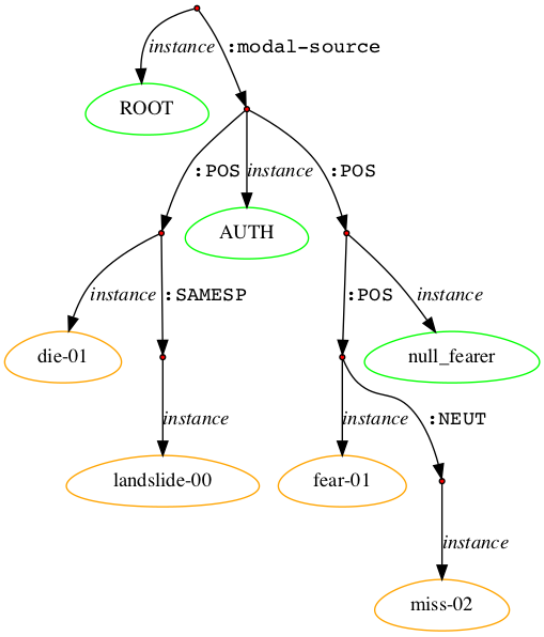
1 Leyte congressman Roger Mercado **said** in a radio **interview** that the village **had** a population of 3,000 to 4,000 and **expressed fears** that as many as 2,000 people **had been buried**.
2 Lerias **called for** more rescuers **to come** to the site **to help look for** bodies as heavy earth moving equipment **could not work** in the mud, which officials **said was** more than six metres (yards) **deep** in many areas.
3 Volunteer rescue teams from the country's mining companies, **skilled** in **digging** through the earth **to rescue** people, **were** also **going** to the area, President Arroyo **said**.
4 Lerias **said** a smaller **landslide** later in the afternoon **caused no damage** but **left** many of the rescuers **worried** about a possible new **disaster**.
5 Relief groups **called for** drinking water, food, blankets and body bags **to be brought** to the scene.

Genre: Narrative

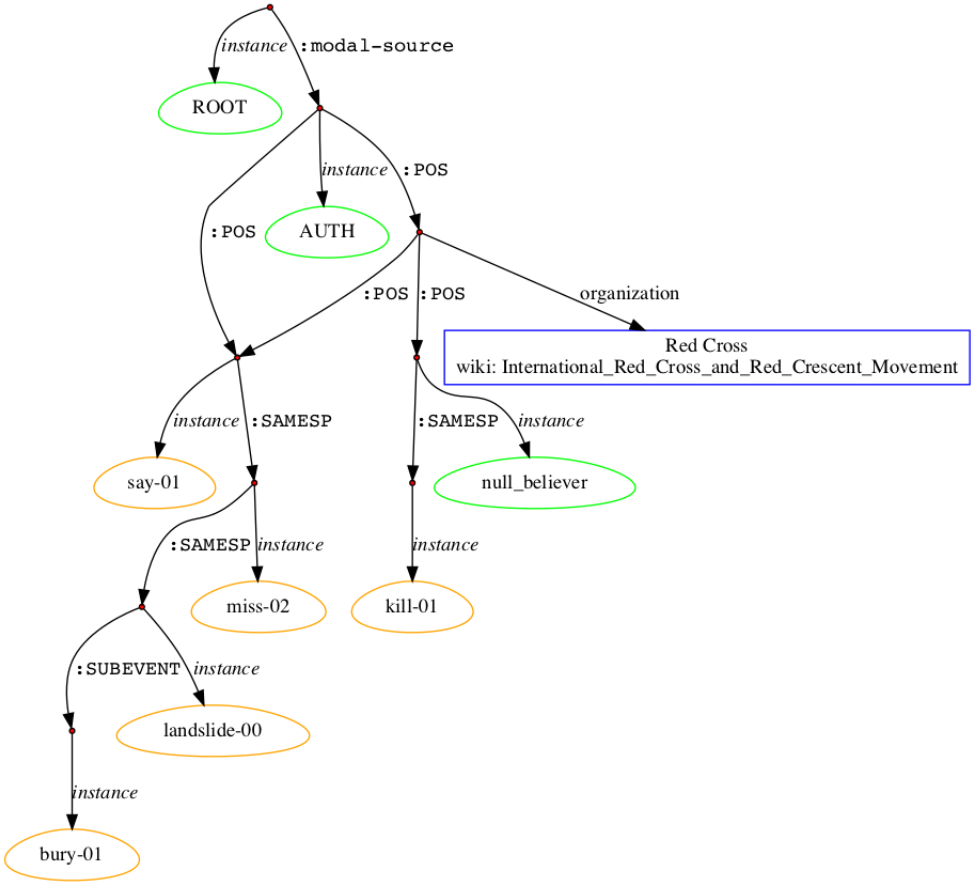


1 My Drawing Number Two **looked** like this: The grown-ups' **response**, this time, was **to advise** me **to lay** aside my drawings of boa constrictors, whether from the inside or the outside, and **devote** myself instead to geography, history, arithmetic and grammar.
2 Grown-ups **never understand** anything by themselves, and it **is tiresome** for children **to be** always and forever **explaining** things to them.
3 So then I **chose** another profession, and **learned to pilot** airplanes.
4 I **have flown** a little over all parts of the world; and it is true that geography **has been** very **useful** to me.
5 At a glance I **can distinguish** China from Arizona.
6 If one **gets lost** in the night, such knowledge **is valuable**.

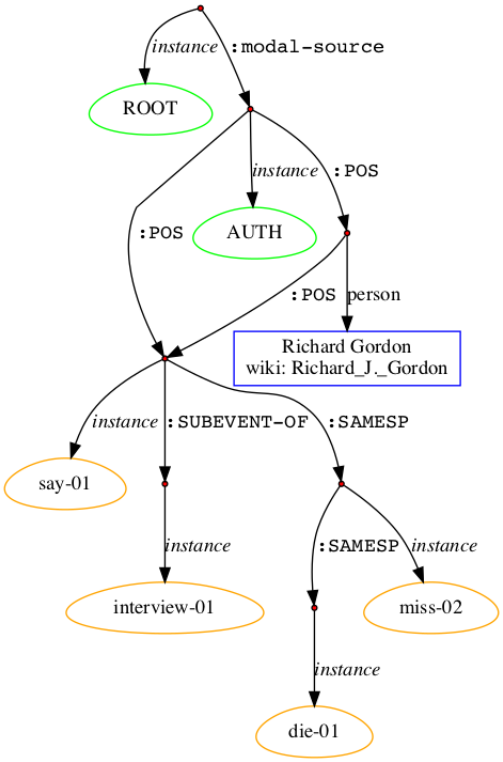
AMR-Reader graphs (Pan et al., 2015)



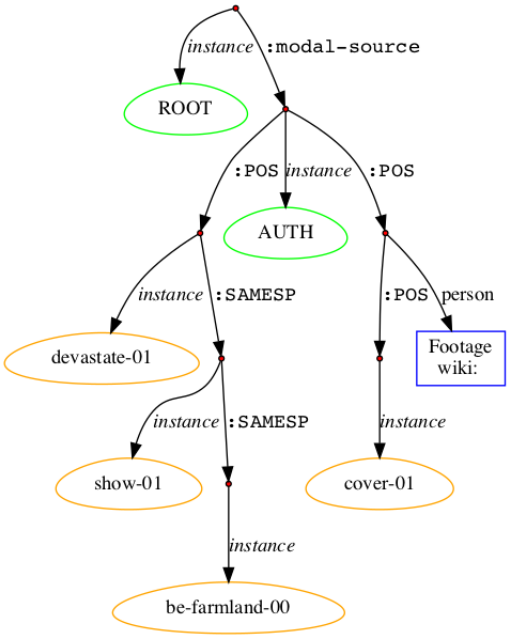
200 dead, 1,500 feared missing in Philippines landslide.



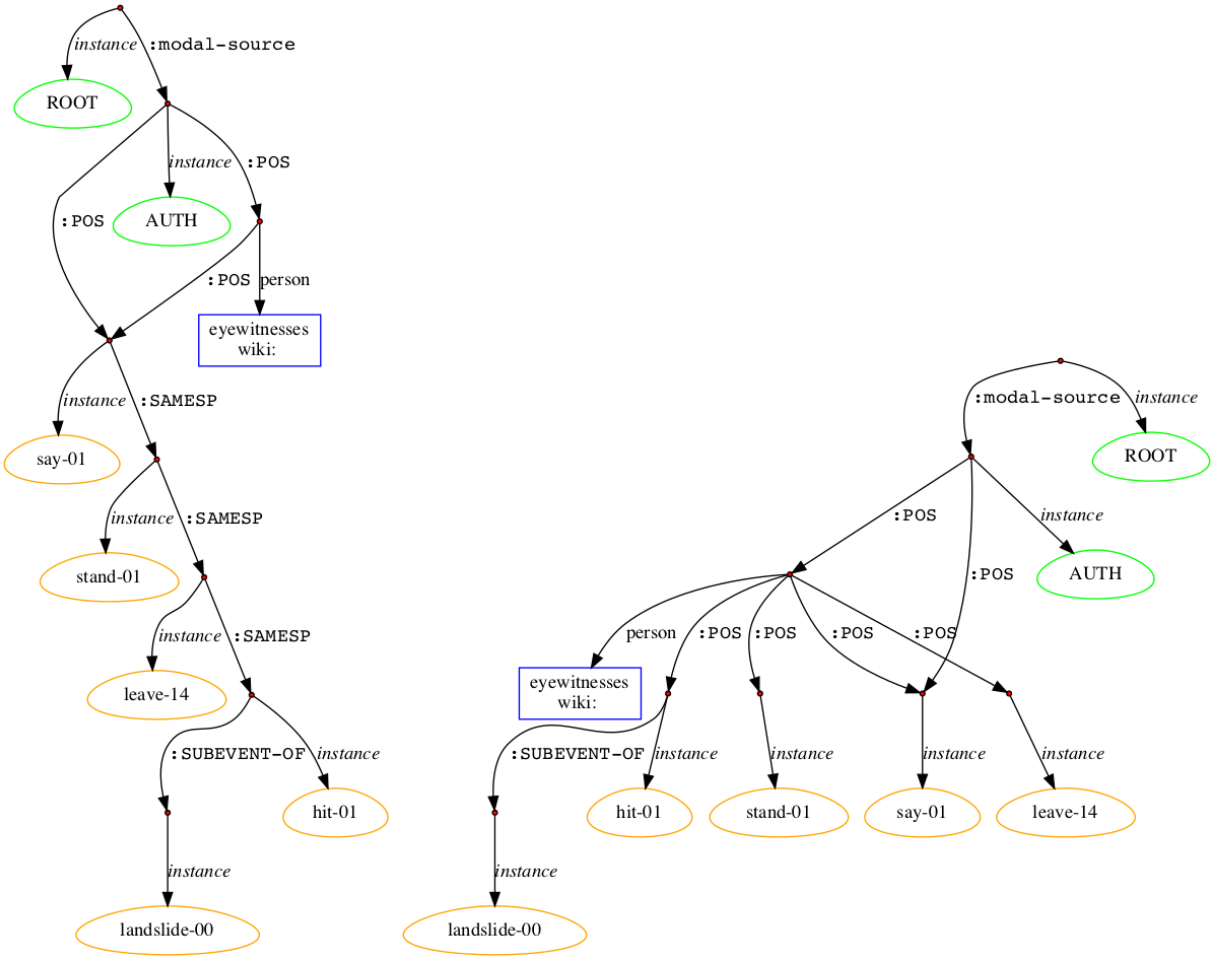
About 200 people were believed killed and 1,500 others were missing in the Central Pilippines on Friday when a landslide buried an entire village, the Red Cross said.



“There are about 1,500 missing, 200 dead,” Richard Gordon, the head of the Philippine Red Cross, said in a radio interview.



The first footage from the devastated village showed a sea of mud covering what had been lush green valley farmland.



Eyewitnesses said only a few houses were left standing after the landslide hit the village of Guinsaugon in the south of the Philippine island of Leyte.

The two graphs on this page show the difference between using the SAMESP annotation (on the left) and not using SAMESP (on the right). We believe that, while SAMESP may introduce too many non-substantive errors into the annotation, it is a useful tool for visualization. This is because it visually groups together events with the same modal strength. Although we have removed the SAMESP edge label in later versions of the annotation scheme, SAMESP may be automatically re-introduced into the annotations for the purpose of visualization.