Toward Fine-grained Annotation of Modality in Text

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz†, and Paul Portner

Department of Linguistics
Georgetown University, Washington, DC
{ar1029,hjh34,eak49,des62,portnerp}@georgetown.edu
†ekatz@caci.com

Abstract

We present a linguistically-informed schema for annotating modal expressions and describe its application to a subset of the MPQA corpus of English texts (Wiebe et al. 2005). The annotation is fine-grained in two respects: (i) in the range of expressions that are defined as modal targets and (ii) in the amount of information that is annotated for each target expression. We use inter-annotator reliability results to support a two-way distinction between priority and non-priority modality types.

1 Introduction

An important part of understanding natural language depends on the ability to tease apart information about the **actual** from that about the **modal**. From the perspective of textual entailment, for example, a non-modal statement like *The butler is the culprit* supports inferences about the actual state of affairs which are quite different from those supported by modal counterparts of the same sentence. Statements of possibility like *It's {possible, somewhat likely} that the butler is the culprit* do not support inferences that are entailed by the non-modal sentence, e.g., that the butler lied to the police when he said he was innocent. Automatically distinguishing between the actual and the modal is necessary for high-quality textual entailment (Burchardt and Frank 2006, Saurí and Pustejovsky 2007, Hickl and Bensley 2007), information extraction (Karttunen and Zaenen 2005), question answering (Saurí et al. 2006), sentiment analysis (Wiebe et al. 2005), and machine translation (Baker et al. 2012). The special attention that modality and non-factuality have received in the context of the textual entailment task is evidence that these aspects of meaning cannot be ignored by practical applications that seek to approximate a complete understanding of text.

From a computational perspective, modally-annotated corpora are an indispensible resource for training systems to automatically interpret modality. This includes automatically detecting modal expressions in text, classifying them into types (as will be described shortly below), identifying their semantic scope, and so on. Theoretical linguists have turned to modal annotation as well, both as a resource for obtaining detailed descriptions of how modality is expressed—within and across languages, and through the historical development of languages—and in order to test the predictions of formal semantic theories of modality (de Haan 2012a, Hacquard and Wellwood 2012, Yanovich 2012).

The annotation of modal meaning is not an easy task. It presents a variety of challenges, relating on the one hand to the practicalities of annotation and on the other to the subtlety of distinctions to be drawn in the modal domain. The

very definition of the set of modal words raises questions about the differences between prototypical exemplars of the class (e.g., *should, can*) and verbs of propositional attitude (e.g., *believe, want*). Modals also tend to be highly ambiguous, with senses that are subtly distinct and overlapping (Kratzer 1981). Their interpretation essentially presents a challenge of Word Sense Disambiguation: *A ten-year-old can drive that truck* can be interpreted both as describing what is sanctioned by law (a deontic use of *can*), and as describing an ability of certain individuals (an ability use of the modal). In a given context, one interpretation (or **modality type**) but not the other may be intended. Mitigating contextual pressures, it has been argued that the syntactic configuration in which a modal appears constrains the range of interpretations it can receive (Cinque 1999, Hacquard 2006). The complex interaction of context, grammar, and lexical content in the expression of modal meaning makes the task of creating high-quality annotated modal corpora particularly important and challenging.

Our aim in the current project is to propose and evaluate a comprehensive and language-independent schema for annotating modality. We experiment with annotating modality types at different levels of granularity, marking up textual spans that describe the **backgrounds** of modal statements (in the sense of Kratzer 1981), as well as comparative and gradable modal expressions. These features are described in detail in Section 3. In the next section, we situate our project in the context of related work.

2 Previous Annotation Efforts

Recent years have seen a number of major efforts of annotating modal expressions in corpora. The majority of these works have targeted English texts (see Table 1 for a snapshot), but there are also notable projects on other languages (de Haan 2012b, Hendrickx et al. 2012). We briefly summarize these representative studies below.

In English, annotation of modality has focused to a large extent just on modal auxiliaries and verbs. Aspects of modal interpretation that have received the most attention concern the proposition in the scope of a modal (Ruppenhofer and Rehbein 2012, Hendrickx et al. for Portuguese), whether or not it occurs in the scope of a negative operator (de Haan 2011, Baker et al. 2012; also much related work on event factuality and sentiment analysis), the source providing the modal background (Ruppenhofer and Rehbein 2012, Hendrickx et al. 2012) and the interpretation of the modal as either epistemic or non-epistemic (i.e., "root"). Hacquard and Wellwood (2012) focus specifically on modals that occur in embedded environments (antecedents of conditionals, questions, and complements of attitude verbs); de Haan (2011, 2012a) correlates the interpretation of modals with register (written versus spoken), grammatical features of their subject (e.g., its person specification), and properties of their verbal complement (e.g., its temporal inflection).

Baker et al.	**Words**	~150 lemmas
	Types	Non-standard (success, effort, intention, ability, belief, and certain root modalities)
	Texts	Written
	# tokens	229
de Haan	**Words**	*must* (and others)
	Types	Root/epistemic
	Texts	Written, spoken
	# tokens	1508 (141)
Hacquard & Wellwood	**Words**	*must, have to*
	Types	Root/epistemic
	Texts	Written
	# tokens	2426
Ruppenhofer & Rehbein	**Words**	*must, ought, shall/should can/could, may/might*
	Types	Epistemic, deontic, dynamic, optative, concessive, conditional
	Texts	Written
	# tokens	1162

Table 1: Recent annotation of modality in English

3 Annotation Task and Proposed Schema

3.1 Overview

With the long-term goal of comparing our annotations to previous work, ultimately producing a resource that is both reliable and built on a platform that is widely used in the field, we followed Ruppenhofer and Rehbein (2012) and chose the MPQA corpus (Wiebe et al. 2005; Wilson 2008) as the corpus to be annotated.

In defining the targets for annotation, we developed a working definition of what counts as modal. This definition was intended to apply not just to modal auxiliaries and verbs (e.g., the typically-targeted *must, have to*), and, moreover, to distinguish modals from closely related attitude verbs. A **modal expression** according to this definition is (i) an expression used to describe alternative ways the world could be, (ii) that has some sort of propositional argument (referred to as the **prejacent**),[1] and (iii) is not associated with an overt attitude holder. Only expressions that met all three criteria were considered modal targets for our annotation. For example, while the noun *hope* is considered a modal in *There is **hope** that she will win*, it is not considered a modal in *There was still **hope*** (condition (ii) is not met). Similarly, *believed* counts as a modal in a sentence like *It is **believed** that…*, but not when it has an attitude holder as its subject (e.g., in *She **believes** that…*).

To make the annotators' task of identifying every modal expression in a text more manageable, a seed list of expressions was generated. The list was used to pre-highlight candidate modal targets in the documents to be annotated. It was compiled according to the procedure in (1) and contained a total of 321 lemmas.

(1) Seed list – lemmas of expressions in (i)-(iv):

i. Adjectives retrieved via the corpus query `[it is [ADJ] that]` in a large corpus (ukWaC, Ferraresi 2007)
ii. Modal expressions annotated in previous work (Table 1)
iii. Synonyms of the expressions in (i), gathered manually using a thesaurus
iv. Modal adverbs, nouns, comparative and superlative adjectives related derivationally to the expressions in (i)-(iii)

Importantly, annotators were instructed to mark every expression that conveyed modality, thereby adding to the pre-highlighted items, and to delete items that were marked in error. Excluded expressions included future *will*, modals in idiomatic phrases (for example, *better* in *the sooner the better*), conditional *should* (as in *Should it be possible, do it!* or *If the negotiations should continue*), and certain teleological verbs (in particular, *aim at/to* in which an attitude holder or goal expression are explicit).

The task was carried out by two annotators. We began with a training round of 40 files, during which the annotators raised issues and discussed problematic cases among themselves and with a larger group of experts. At the end of the training phase the annotation guidelines had been finalized. In total, the annotators completed 200 files (183 of which contained at least one modal target according to at least one annotator). These included 1232 fully annotated targets that both annotators agreed were modal and that had identical spans. There was considerable disagreement concerning what counts as modal, as can be seen in Table 2.

Annotator 1	1605
Annotator 2	1810
Total agreed modal	1232

Table 2: Number of tokens annotated

3.2 Tool

We used the MMAX2 tool (Müller and Strube 2006) for our annotation, following Hendrickx et al.'s positive experience. In addition to its relatively user-friendly graphical interface, MMAX2 can accommodate annotation of overlapping and discontinuous elements. As noted by these authors, these abilities are crucial for annotation of modal features (see (7) below for an example involving a discontinuous prejacent).

[1] An exception are adnominal modal adjectives. See discussion of Modified Elements below.

Figure 1 shows a target modal as it appears to the annotator in MMAX2. Connecting lines are drawn between the modal and spans of text that correspond to its prejacent, and other features potentially represented in the text.

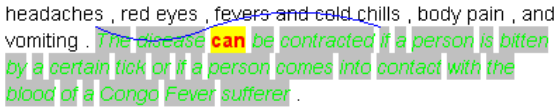

Figure 1: Modal target *can* annotated in MMAX2. One line connects the modal to its Prejacent, and a second one connects it to its Background.

3.3 Annotated Features

This section presents our annotation schema. We discuss the central features of the schema, focusing on those that are new to this project: Modality Type (coarse-grained and fine-grained), Propositional Arguments (including prejacents and comparatives), Background, Modified Element, Degree Indicator, and Outscoping Quantifier. More detailed descriptions of all features appear in the annotation guidelines, which will be made available separately.[2]

Modality Type. Every modal was categorized on two levels with respect to the type of modality it conveyed in context. Seven fine-grained types were distinguished: Epistemic, Circumstantial, Ability, Deontic, Bouletic, Teleological, and Bouletic/Teleological. However, before this classification was made, annotators first categorized each modal as belonging to one of three coarse-grained categories: Epistemic or Circumstantial, Ability or Circumstantial, and Priority. The label **priority** picks out a conceptually motivated subclass of non-epistemic modalities: those that use some "priority" (a desire, a goal) to designate certain possibilities as better than others (Portner 2009:135ff.)[3] In Section 4 we show that annotators reliably agreed on only the highest level split between priority and non-priority interpretations.

[2] Examples from the ukWaC corpus are noted below as [Web], and ones from the annotated corpus are noted as [MPQA].

[3] A special category of "TBD" (to-be-discussed) was available for annotators to mark unclear examples that should be revisited. This category was only used a handful of times.

Non-priority

- *Epistemic*: the claim is based on evidence, belief or knowledge.
- *Circumstantial*: the claim is based on circumstances.
- *Ability:* the claim is based on what someone/something can do.

(2)

a. Mary ***must*** have a good reason for being late. (Epistemic)
b. … for we were in the little salon where Madame never sat in the evening, and where it was by mere ***chance*** that heat was still lingering in the stove. [Web] (Circumstantial)
c. The potential losses that could be incurred by the tourist industry following a major disaster ***can*** be illustrated by examining the consequences of hurricanes Luis and Marilyn to the Caribbean island of Anguilla in 1995. [Web] (Ability)

In cases of ambiguity, annotators were instructed to mark the modality type on the fine-grained level as Ability/Circumstantial (interpreted as both ability and circumstantial) or Epistemic/Circumstantial, as appropriate. Another special use of the Ability/Circumstantial label was reserved for opportunity modals (for example, the interpretation of *can* in *You can see the ocean through this window*).

(3)

a. "…but I say: Please, that is the most dangerous thing you ***can*** ever do," he said. [MPQA] (Ability/Circumstantial)
b. Temperatures are very ***likely*** to be significantly higher when in full screen mode because your graphics card will be running in 3D mode. [Web] (Epistemic/Circumstantial)

Priority

- *Deontic*: the claim is based on rules, standards, (social) norms.
- *Bouletic:* the claim is based on someone's wishes or desires.
- *Teleological:* the claim is goal-oriented.
- *Bouletic/Teleological*: for tokens that are arguably both bouletic and teleological.

(4)

a. The rich ***must*** give money to the poor. (Deontic)
b. Today, he is being completely isolated and the ***desire*** to drive him away is scarcely disguised. [MPQA] (Bouletic)
c. The owner and a neighbor who had helped him put down the animal were sent ***urgently*** to the Hospital for Infectious Diseases in Miercurea Ciuc, where they received preventive anti-rabies treatment. [MPQA] (Teleological)
d. The donors' conference, […], was hoping to raise 1.25 billion dollars (1.47 billion euros) for Yugoslavia this year for ***urgent*** repairs to infrastructure and salaries to teachers and other civil servants. [MPQA] (Bouletic/Teleological)

In cases of an ambiguity between deontic and any other priority-type modality, a modal was given the Priority subtype (the same label as its coarse-grained classification).

(5)

When it gets to the point that Northern Alliance troops start firing in the air just next to a car with reporters, you ***have to*** do something about it, " said Cordell. [MPQA] (ambiguous Priority)

In addition to modality type, every target modal was also specified for its environment's polarity and associated with a prejacent in the text.

Environmental Polarity. The environment of the modal was set to 'positive' by default. In cases where the modal was in a semantically downward entailing environment, the value of the feature was changed to 'negative' and the item creating the negative environment (e.g., *not* or *reject*) marked in the text. In cases where modals were in the scope of multiple negative words, these were all marked.

(6)

"There ***could*** be no expediency, no compromise, no lapse in vigilance," he said." [MPQA]

Note that the environment of the modal was not affected by the modal's internal polarity: an inherently negative modal such as *unlikely* does not create a negative environment. Also, not every combination of a modal with negation results in a negative environmental polarity for the modal (e.g. *should* in *should not* outscopes the negation).

Propositional Arguments. The textual span corresponding to the proposition a modal applies to was annotated as the modal's prejacent. Prejacents excluded non-restrictive relative clauses and parentheticals, tense markers (*is* in (7)), expletive *it* (see (7)), markers of environmental polarity, and degree indicators (see below).

(7)

It is ***likely*** that John, who was my upstairs neighbor, will run the race.

Determining whether temporal adverbials fell within the prejacent of the modal was left to annotator discretion.

(8)

According to military experts, it is ***possible*** that clashes will resume between the Taleban and UIFSA forces in various regions of Afghanistan in the next few days and weeks. [MPQA]

As modals may also be used comparatively, annotators could mark when a modal appeared in equative, comparative, or superlative forms. In these instances, prejacents as well as *than*-clauses in the text were associated with the modal.

(9)

a. It is ***likelier*** that John will run a race.
(Comparative; the prejacent is underlined and there is no *than*-clause)
b. John is as ***likely*** to run a race as HE is TO CLIMB A MOUNTAIN.
(Equitive; Prejacent and than-clause underlined, the latter in small caps)

The remaining features are independent of the modal, thus they were only marked if they appeared in the text.

Source. This feature was designed to indicate the entity that had the ability or knowledge that are the basis for a modal claim, or in the case of priority modals, the entity placing the obligation

or setting the goal that the modal takes into account. (This definition of Source is similar to the one proposed by Hendrickx et al. 2012 and Ruppenhofer and Rehbein 2012). Annotators marked the closest instance of reference to the source, pronominal or otherwise. Sources could be inanimate.

(10)
In his latest speech, Chen said the long-standing dispute with China ***must*** be resolved through dialogue with respect to the principles of democracy and freedom. [MPQA] (Deontic Source)

Where two or more possible sources were detected, as in (11), no Source was marked.[4]

(11)
Chang said after visiting Chinese communities in the United States, New Zealand and Southeast Asia that education and cultural work ***needs*** to be further strengthened in the Chinese communities in Southeast Asia. [MPQA]

Background. The background of a modal is a sequence of (one or more) constituents that provide a textual description of the circumstances and/or priorities that the modal claim is based on.[5] The background may be expressed in an adjunct that contains a description of a relevant situation; in the case of a priority modal, (12), a rationale clause may describe the relevant goals and preferences.

(12)
With the new method, all you ***need*** do to get an answer is put all the ingredients into a test tube, mix them together, and check to see what the output is. [MPQA]

Modified Element. We also included in our annotation modals that were not used predicatively, but as modifiers of nouns or adjectives (as in, *the* ***probable*** *answer*, *It was* ***sufficiently*** concrete). The head (underlined in the examples above) of the modified phrase was marked as a Modified Element and associated with the modal in these cases.[6]

Degree Indicator. Any item that indicated degrees of modal necessity or possibility was annotated. In cases where two or more degree indicators were used, they were treated as one degree indicator for purposes of the annotation.

(13)
There is a very high ***likelihood*** that it will rain.

Adverbs like *perhaps* were not treated as degree indicators, but as independent modals (in cases such as *It could perhaps be...*).

Outscoping Quantifier. Quantificational elements that are part of a modal's prejacent but are nevertheless interpreted as taking semantic scope over the modal were marked as outscoping quantifiers.

(14)
Everyone ***can*** win the prize.

Finally, additional features for each modal were its **Lemma** (included automatically for pre-highlighted modals), and a text box for optional comments (used, e.g., for indicating that the modal was in the title of the document). Table 3 summarizes the features annotated and their possible values.

Feature	**Possible Values**
Modality type (coarse)	Epistemic or Circumstantial, Ability or Circumstantial, Priority
Modality type (fine)	Epistemic, Circumstantial, Ability, Deontic, Bouletic, Teleological, Bouletic/Teleological
Environmental polarity	Positive, Negative
Propositional arguments	textual span(s)
Source	textual span

[4] An alternative strategy would be to mark multiple Sources in such cases, or to mark the more plausible source (in this example, *Chang*).

[5] A related feature tracking whether there is "overt evidence" for a *must* claim in the text is raised by de Haan (2011).

[6] The phrase modified by the modal (e.g., *repairs to infrastructure* in (4d)) was marked as the modal's prejacent.

Background	textual span
Modified element	textual span
Degree indicator	textual span
Outscoping quantifier	textual span
Lemma	free text
Additional notes	free text

Table 3: Annotated features and their values

4 Reliability of the Annotations

As an indication of the reliability of our annotation, we measured inter-annotator agreement on the following features: Modality Type, Prejacent, and Background. The results obtained allow for initial comparison with previous annotation projects.

4.1 Inter-annotator Reliability Measure: Krippendorff's α

Inter-annotator reliability is standardly measured with a number of different α and κ scores (Carletta 1996, Artstein and Poesio 2008). These scores measure the degree of agreement for a pair or set of annotators given some common set of annotation guidelines. Though the details of calculation and presupposed conditions vary among different scores, the statistics themselves are comparable, with 0.0 reflecting agreement no better than would occur at chance and 1.0 reflecting perfect agreement.

Many reliability measures presuppose identical annotation items. In our case, this is insufficient, as the annotators are marking features with values that are spans of text of potentially variable position and length (e.g., Prejacent, Background). For this reason, Krippendorff's (2004) α score was selected as a measure of inter-annotator reliability. The agreement score is computed along a continuum, comparing the overlap between spans that were marked by the annotators, and allowing for partial agreement when spans are not perfectly aligned.

In our analysis, we included only the 1232 agreed-upon modals (targets that were marked as modal by both annotators and that had identical spans). We ignored cases of partial overlap (where one annotator marked *would like* and the other its substring *would*, for example). We calculated inter-annotator agreement on Modality Type, Prejacent (to the exclusion of other propositional arguments), and Background.

4.2 Results

Table 4 shows the reliability of annotations for the three features mentioned above. We begin with discussion of Modality Type, the feature that has received most attention in previous work. The basic score, measuring agreement on the ten possible values for this feature, was low ($\alpha = 0.49$). The effect of a number of category collapses was therefore investigated. First, all priority-type categories were collapsed (i.e., Bouletic, Teleological, Bouletic/Teleological, Deontic, and Priority were treated as one category for purposes of the agreement calculation; "Priority types collapsed" in Table 4). Second, all non-priority types (i.e., Epistemic, Circumstantial, Ability, Epistemic/Circumstantial, Ability/Circumstantial) were collapsed into one, and finally, both of these collapses were applied together. This final merging of categories (the final row in the Table) resulted in a high α score of 0.89.

Conventionally, κ and α scores are considered acceptable at a threshold of 0.80 (Carletta 1996). A word sense disambiguation task (Ng et al. 1999) has been reported with raw κ scores at 0.317, and 0.862 after using a collapsing algorithm to attain better agreement. Hacquard and Wellwood (2012) achieved a κ score of 0.84 on a two-way classification of epistemic versus root modalities. Table 5 compares our α agreement scores for individual modals with the κ scores reported on a superset of the data by Ruppenhofer and Rehbein ([RR 2012] below).

Feature: Modality Type	α
No collapse	0.49
Priority types collapsed	0.66
Non-priority types collapsed	0.73
Priority vs. Non-priority	0.89
Feature: Prejacent	0.65
Feature: Background	0.61

Table 4: Inter-annotator reliability scores

	Items	Raw α	Non-P collapsed	P collapsed	Priority vs Non-priority	[RR 2012] κ
may, might	102	0.27	1.00	N/A	1.00	0.621
must	140	0.40	0.40	0.90	0.90	0.848
shall, should[7]	140	0.23	0.31	0.71	0.80	0.602
can	238	0.34	0.90	0.35	0.91	0.614

Table 5: Modality Type agreement by word

Agreement scores for the Prejacent and Background features, α = 0.65 and α = 0.61 respectively, represent how well the spans that annotators marked align overall (although note that they do not take into account the association of particular spans to modals in the text). Prejacents have been annotated in previous studies, but inter-annotator agreement scores for this feature have not previously been reported, as far as we know.

4.3 Discussion

Collapsing the priority types as well as the non-priority types results in essentially a two-way distinction that is similar to the Root versus Epistemic distinction assumed in previous annotation projects (see Table 1). Our results support making the distinction at this coarse-grained level. Since the α score is designed to cancel out any random (dis)agreement, the increasing values with each collapse show that the annotators cannot reliably discern the more fine-grained distinctions of modal "flavor". (Rubinstein et al. 2012 report similar results in a crowdsourcing experiment with non-expert native speakers.) Nevertheless, while the coarse-grained distinctions may prove more reliable for Machine Learning and other NLP applications, having access to the annotators' finer-grained judgments could be helpful for theoretical purposes. They allow the researcher to distinguish between more and less ambiguous exemplars of each modality type and to investigate the grammatical and contextual properties of examples that are judged as more ambiguous. We thus propose to annotate for each modal a unique coarse-grained type as well as a list of one or more fine-grained types corresponding to the annotators' individual judgments.

Since collapsing non-priority subtypes results in a greater increase to the 0.49 α baseline (0.73 vs. 0.66 for priority collapsed), we conclude that the difficulty to distinguish epistemic, circumstantial, and ability modalities is greater than the difficulty to distinguish between different subtypes of priority modality. In the confusion matrix below, the affinity between the ability and circumstantial types is evident in the decisions of row annotator within the **a**-column, and in the decisions of column annotator across the **c**-row.

	a	a+c	e+c	e	c	b	b+t	t	d	p
a	55	4	1	7	6	0	1	0	0	0
a+c	2	1	0	0	1	0	0	0	0	0
e+c	0	0	0	6	2	0	0	0	1	0
e	1	1	8	139	30	0	0	1	10	0
c	87	23	20	110	176	0	0	12	18	3
b	0	0	0	2	2	4	2	1	1	2
b+t	0	0	0	5	3	9	1	2	4	5
t	11	2	1	16	27	4	4	102	76	35
d	11	1	0	7	2	0	0	3	126	3
p	1	1	0	0	0	0	1	1	24	4

Table 6: Modality Type confusion matrix (**a**-ability, **c**-circumstantial, **b**-bouletic, **d**-deontic, **e**-epistemic, **p**-priority, **t**-teleological, **a+c**-ability and circumstantial, similarly **e+c** and **b+t**)

Cells marked in red in the confusion matrix evidence confusion between teleological and deontic interpretations on the one hand, and epistemic/circumstantial and even ability interpretations on the other. We leave investigation of the relevant examples to future work.

5 Conclusion

We have proposed a schema for annotating modal meaning that builds on previous work and extends it with a number of novel features. Completing the annotation of the MPQA corpus according to this schema, we hope to contribute to the development of reliable computational resources for detecting and interpreting modals in naturally occurring text.

Acknowledgments

This research was funded by National Science Foundation grant BCS-1053038, "The Semantics of Gradable Modal Expressions", to Graham Katz, Paul Portner, and Elena Herburger.

[7] No instance of *shall* was tagged by both our annotators.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. Extended version of an article in *Computational Linguistics*.

Kathryn Baker, 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 MT. *Computational Linguistics* 38(2):1-28.

Aljoscha Burchardt and Anette Frank. 2006. Approaching textual entailment with LFG and FrameNet frames. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment (RTE-2).*

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2): 249-245.

Guglielmo Cinque. 1999. *Adverbs and functional heads*. Oxford University Press.

Adriano Ferraresi. 2007. Building a very large corpus of English obtained by Web crawling: ukWaC. MA Thesis, University of Bologna.

Ferdinand de Haan. 2011. Disambiguating modals: constructions and *must*. Ms.

Ferdinand de Haan. 2012a. The relevance of constructions for the interpretation of modal meaning: The case of *must*. *English Studies* 93(6): 700-728.

Ferdinand de Haan. 2012b. Automatic disambiguation of modals and evidentials: A corpus-based investigation. Slides of talk presented at the conference "The Nature of Evidentiality", Leiden University, June.

Valentine Hacquard. 2006. Aspects of modality. Doctoral Dissertation, MIT.

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

Iris Hendrickx, Amália Mendes and Silvia Mencarelli. 2012. Modality in text: A proposal for corpus annotation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*

Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the workshop on textual entailment and paraphrasing*.

Lauri Karttunen and Annie Zaenen. 2005. Veridicity. In *Dagstuhl seminar proceedings*. *Annotating, extracting and reasoning about time and events*, ed. G. Katz, J. Pustejovsky, and F. Schilder.

Angelika Kratzer. 1981. The notional category of modality. In *Words, worlds, and contexts*, ed. H.-J. Eikmeyer, H. Rieser, 38-74. Walter de Gruyter.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. Annenberg School for Communication Department Papers, University of Pennsylvania.

Hwee Tou Ng, Chung Yong Lim, and Shou King Foo. 1999. A case study on inter-annotator agreement for Word Sense Disambiguation. In *Proceedings of the Siglex-ACL Workshop on Standarizing Lexical Resources*.

Christoph Müller and Michael Strube. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, ed. S. Braun, K. Kohn, J. Mukherjee, 197-214. Peter Lang.

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

Aynat Rubinstein, Dan Simonson, Joo Chung, Hillary Harner, Graham Katz, and Paul Portner. 2012. Developing a methodology for modality type annotations on a large scale. Slides of talk presented at the Modality Workshop, Ottawa University, April.

Josep Ruppenhofer and Ines Rehbein. 2012. Yes we can!? Annotating English modal verbs. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Roser Saurí and James Pustejovsky. 2007. Determining modality and factuality for textual entailment. In *ICSC-07*.

Roser Saurí, Marc Verhagen and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *19th International FLAIRS conference (FLAIRS 2006)*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 37(2-3):165-210.

Theresa Wilson. 2008. Fine-grained subjectivity analysis. Doctoral Dissertation, University of Pittsburgh.

Igor Yanovich. 2012. Possibility or collapse on the path to necessity? Slides of talk presented at Georgetown University, November.