Did It Happen? The Pragmatic Complexity of Veridicality Assessment

Marie-Catherine de Marneffe[*]
Stanford University

Christopher D. Manning[**]
Stanford University

Christopher Potts[†]
Stanford University

Natural language understanding depends heavily on assessing veridicality*—whether events mentioned in a text are viewed as happening or not—but little consideration is given to this property in current relation and event extraction systems. Furthermore, the work that has been done has generally assumed that veridicality can be captured by lexical semantic properties whereas we show that context and world knowledge play a significant role in shaping veridicality. We extend the FactBank corpus, which contains semantically driven veridicality annotations, with pragmatically informed ones. Our annotations are more complex than the lexical assumption predicts but systematic enough to be included in computational work on textual understanding. They also indicate that veridicality judgments are not always categorical, and should therefore be modeled as distributions. We build a classifier to automatically assign event veridicality distributions based on our new annotations. The classifier relies not only on lexical features like hedges or negations, but also on structural features and approximations of world knowledge, thereby providing a nuanced picture of the diverse factors that shape veridicality.*

"All I know is what I read in the papers"
—Will Rogers

1. Introduction

A reader's or listener's understanding of an utterance depends heavily on assessing the extent to which the speaker (author) intends to convey that the events described did (or did not) happen. An unadorned declarative like *The cancer has spread* conveys firm speaker commitment, whereas qualified variants such as *There are strong indicators that the cancer has spread* or *The cancer might have spread* imbue the claim with uncertainty. We

[*] Linguistics Department, Margaret Jacks Hall Building 460, Stanford CA 94305, USA.
E-mail: mcdm@stanford.edu.
[**] Linguistics Department & Computer Science Department, Gates Building 1A, 353 Serra Mall, Stanford CA 94305, USA. E-mail: manning@stanford.edu.
[†] Linguistics Department, Margaret Jacks Hall Building 460, Stanford CA 94305, USA.
E-mail: cgpotts@stanford.edu.

Submission received: 4 April 2011; revised submission received: 1 October 2011; accepted for publication: 30 November 2011.

call this **event veridicality**, building on logical, linguistic, and computational insights about the relationship between language and reader commitment (Montague 1969; Barwise 1981; Giannakidou 1994, 1995, 1999, 2001; Zwarts 1995; Asher and Lascarides 2003; Karttunen and Zaenen 2005; Rubin, Liddy, and Kando 2005; Rubin 2007; Saurí 2008). The central goal of this article is to begin to identify the linguistic and contextual factors that shape readers' veridicality judgments.[1]

There is a long tradition of tracing veridicality to fixed properties of lexical items (Kiparsky and Kiparsky 1970; Karttunen 1973). On this view, a lexical item L is **veridical** if the meaning of L applied to argument p entails the truth of p. For example, because both true and false things can be believed, one should not infer directly from *A believes that S* that S is true, making *believe* non-veridical. Conversely, *realize* appears to be veridical, because realizing S entails the truth of S. The prototypical anti-veridical operator is negation, because *not S* entails the falsity of S, but anti-veridicality is a characteristic of a wide range of words and constructions (e.g., *have yet to*, *fail*, *without*). These basic veridicality judgments can be further subdivided using modal or probabilistic notions. For example, although *may* is non-veridical by the basic classifications, we might classify *may S* as *possible* with regard to S.[2]

Lexical theories of this sort provide a basis for characterizing readers' veridicality judgments, but they do not tell the whole story, because they neglect the pragmatic enrichment that is pervasive in human communication. In the lexical view, *say* can only be classified as *non-veridical* (both true and false things can be said), and yet, if a *New York Times* article contained the sentence *United Widget said that its chairman resigned*, readers would reliably infer that United Widget's chairman resigned—the sentence is, in this context, veridical (at least to some degree) with respect to the event described by the embedded clause, with *United Widget said* functioning to mark the source of evidence (Simons 2007). *Cognitive authority*, as termed in information science (Rieh 2010), plays a crucial role in how people judge the veridicality of events. Here, the provenance of the document (the *New York Times*) and the source (United Widget) combine to reliably lead a reader to infer that the author intended to convey that the event really happened. Conversely, *allege* is lexically non-veridical, and yet this only begins to address the complex interplay of world knowledge and lexical meaning that will shape people's inferences about the sentence *FBI agents alleged in court documents today that Zazi had admitted receiving weapons and explosives training from al Qaeda operatives in Pakistan last year*.

We conclude from examples like this that veridicality judgments have an important pragmatic component, and, in turn, that veridicality should be assessed using information from the entire sentence as well as from the context. Lexical theories have a significant role to play here, but we expect their classifications to be buffeted by other communicative pressures. For example, the lexical theory can tell us that, as a narrowly semantic fact, *X alleges S* is non-veridical with regard to S. Where X is a trustworthy source for S-type information, however, we might fairly confidently conclude that S is true. Where X is known to spread disinformation, we might tentatively conclude that S is false. These pragmatic enrichments move us from uncertainty to some degree of

[1] Our use of the term "veridicality" most closely matches that of Giannakidou (1999), where it is defined so as to be (i) relativized to particular agents or perspectives, (ii) gradable, and (iii) general enough to cover not only facts but also the commitments that arise from using certain referential expressions and aspectual morphology. The more familiar term "factuality" seems at odds with all three of these criteria, so we avoid it.

[2] Lexical notions of veridicality must be relativized to specific argument positions, with the other arguments existentially closed for the purposes of checking entailment. For example, *believe* is non-veridical on its inner (sentential) argument because "$\exists x : x$ believes p" does not entail p.

certainty. Such enrichments can be central to understanding how a listener (or a reader) understands a speaker's (or author's) message.

Embracing pragmatic enrichment means embracing uncertainty. Although listeners can feel reasonably confident about the core lexical semantics of the words of their language, there is no such firm foundation when it comes to this kind of pragmatic enrichment. The newspaper says, *United Widget said that its profits were up in the fourth quarter*, but just how trustworthy is United Widget on such matters? Speakers are likely to vary in what they intend in such cases, and listeners are thus forced to operate under uncertainty when making the requisite inferences. Lexical theories allow us to abstract away from these challenges, but a pragmatically informed approach must embrace them.

The FactBank corpus is a leading resource for research on veridicality (Saurí and Pustejovsky 2009). Its annotations are "textual-based": They seek to capture the ways in which lexical meanings and local semantic interactions determine veridicality judgments. In order to better understand the role of pragmatic enrichment, we had a large group of linguistically naive annotators annotate a portion of the FactBank corpus, given very loose guidelines. Whereas the FactBank annotators were explicitly told to avoid bringing world knowledge to bear on the task, our annotators were encouraged to choose labels that reflected their own natural reading of the texts. Each sentence was annotated by 10 annotators, which increases our confidence in them and also highlights the sort of vagueness and ambiguity that can affect veridicality. These new annotations help confirm our hypothesis that veridicality judgments are shaped by a variety of other linguistic and contextual factors beyond lexical meanings.

The nature of such cues is central to linguistic pragmatics and fundamental to a range of natural language processing (NLP) tasks, including information extraction, opinion detection, and textual entailment. Veridicality is prominent in BioNLP, where identifying negations (Chapman et al. 2001; Elkin et al. 2005; Huang and Lowe 2007; Pyysalo et al. 2007; Morante and Daelemans 2009) and hedges or "speculations" (Szarvas et al. 2008; Kim et al. 2009) is crucial to proper textual understanding. Recently, more attention has been devoted to veridicality within NLP, with the 2010 workshop on negation and speculation in natural language processing (Morante and Sporleder 2010). Veridicality was also at the heart of the 2010 CoNLL shared task (Farkas et al. 2010), where the goal was to distinguish uncertain events from the rest. The centrality of veridicality assessment to tasks like event and relation extraction is arguably still not fully appreciated, however. At present the vast majority of information extraction systems work at roughly the clause level and regard any relation they find as true. But relations in actual text may not be facts for all sorts of reasons, such as being embedded under an attitude verb like *doubt*, being the antecedent of a conditional, or being part of the report by an untrustworthy source. To avoid wrong extractions in these cases, it is essential for NLP systems to assess the veridicality of extracted facts.

In the present article, we argue for three main claims about veridicality. First and foremost, we aim to show that pragmatically informed veridicality judgments are systematic enough to be included in computational work on textual understanding. Second, we seek to justify FactBank's seven-point categorization over simpler alternatives (e.g., certain vs. uncertain, as in the CoNLL task). Finally, the inherent uncertainty of pragmatic inference suggests to us that veridicality judgments are not always categorical, and thus are better modeled as probability distributions over veridicality categories. To substantiate these claims, we analyze in detail the annotations we collected, and we report on experiments that treat veridicality as a distribution-prediction task. Our feature set includes not only lexical items like hedges, modals, and negations, but also complex structural features and approximations of world knowledge. Though the resulting classifier has limited ability to assess veridicality in complex real world contexts, it still does a quite good job of capturing human pragmatic judgments of veridicality. We argue that the model yields insights into the complex pragmatic factors that shape readers' veridicality judgments.

2. Corpus Annotation

FactBank's annotations are intended to isolate semantic effects from pragmatic ones in the area of veridicality assessment. Our overarching goal is to examine how pragmatic enrichment affects this picture. Thus, we use the FactBank sentences in our own investigation, to facilitate comparisons between the two kinds of information and to create a supplement of FactBank itself. This section introduces FactBank in more detail and then thoroughly reviews our own annotation project and its results.

2.1 FactBank Corpus

FactBank provides veridicality annotations on events relative to each participant involved in the discourse. It consists of 208 documents from newswire and broadcast news reports in which 9,472 event descriptions (verbs, nouns, and adjectives) were manually identified. There is no fundamental difference in the way verbs, nouns, and adjectives are annotated. Events are single words. The data come from TimeBank 1.2 and a fragment of AQUAINT TimeML (Pustejovsky et al. 2006). The documents in the AQUAINT TimeML subset come from two topics: "NATO, Poland, Czech Republic, Hungary" and "the Slepian abortion murder."

The tags annotate ⟨event, participant⟩ pairs in sentences. The participant can be anyone mentioned in the sentence as well as its author. In Example (1), the target event identified by the word *means* is assigned a value with respect to both the source *some experts* and the author of the sentence.

Example 1
Some experts now predict Anheuser's entry into the fray **means** near-term earnings trouble for all the industry players.

Veridicality(means, experts) = PR+

Veridicality(means, author) = Uu

The tags are summarized in Table 1. Each tag consists of a veridicality value (certain [CT], probable [PR], possible [PS], underspecified [U]) and a polarity value (positive [+], negative [−], unknown [u]). CT+ corresponds to the standard notion of veridicality, CT− to anti-veridicality, and Uu to non-veridicality. The PR and PS categories add a modal or probabilistic element to the scale, to capture finer-grained intuitions.

Examples (2) and (3) illustrate the annotations for a noun and an adjective.

Example 2
But an all-out bidding **war** between the world's top auto giants for Britain's leading luxury-car maker seems unlikely.

Veridicality(war, author) = PR−

Table 1
FactBank annotation scheme. CT = certain; PR = probable; PS = possible; U = underspecified; + = positive; − = negative; u = unknown.

Value	Definition	Count
CT+	According to the source, it is certainly the case that X	7,749 (57.6%)
PR+	According to the source, it is probably the case that X	363 (2.7%)
PS+	According to the source, it is possibly the case that X	226 (1.7%)
CT−	According to the source, it is certainly not the case that X	433 (3.2%)
PR−	According to the source it is probably not the case that X	56 (0.4%)
PS−	According to the source it is possibly not the case that X	14 (0.1%)
CTu	The source knows whether it is the case that X or that not X	12 (0.1%)
Uu	The source does not know what the factual status of the event is, or does not commit to it	4,607 (34.2%)
		13,460

Example 3
Recently, analysts have said Sun also is **vulnerable** to competition from International Business Machines Corp., which plans to introduce a group of workstations early next year, and Next Inc.

> Veridicality(vulnerable, analysts) = CT+
>
> Veridicality(vulnerable, author) = Uu

The last column of Table 1 reports the value counts in the corpus. The data are heavily skewed, with 62% of the events falling to the positive side and 57.6% in the CT+ category alone. The inter-annotator agreement for assigning veridicality tags was high ($\kappa = 0.81$, a conservative figure given the partial ordering in the tags). A training/test split is defined in FactBank: The documents from TimeBank 1.2 are used as the training data and the ones from the subset of the AQUAINT TimeML corpus as the testbed.

As we noted earlier, FactBank annotations are supposed to be as purely semantic as possible; the goal of the project was to "focus on identifying what are the judgments that the relevant participants make about the factuality nature of events, independently from their intentions and beliefs, and exclusively based on the linguistic expressions employed in the text to express such judgments," disregarding "external factors such as source reliability or reader bias" (Saurí 2008, page 5). The annotation manual contains an extensive set of discriminatory tests (Saurí 2008, pages 230–235) that are informed by lexical theories of veridicality. The resulting annotations are "textual-based, that is, reflecting only what is expressed in the text and avoiding any judgment based on individual knowledge" (Saurí and Pustejovsky 2009, page 253). In addition, discourse structure is not taken into account: "we decided to constrain our annotation to information only present at the sentence level" (Saurí and Pustejovsky 2009, page 253).

2.2 Annotations from the Reader's Perspective

In the terminology of Levinson (1995, 2000), FactBank seeks to capture aspects of *sentence meaning*, whereas we aim to capture aspects of *utterance meaning*, which brings us closer to characterizing the amount and kind of information that a reader can reliably extract from an utterance. We thus extend the FactBank annotations by bringing world knowledge into the picture. Whereas the FactBank annotators were explicitly told to avoid any reader bias, to disregard the credibility of the source, and to focus only on the linguistic terms used in the text to express veridicality, we are interested in capturing how people judge the veridicality of events when reader bias, credibility of the source, and what we know about the world is allowed to play a role.

To do this, we took a subset of the FactBank sentences annotated at the author level and recruited annotators for them using Mechanical Turk. We restricted the task to annotators located in the United States. Our subset consists of 642 sentences (466 verbs, 155 nouns, 21 adjectives); we use all the PR+, PS+, PR−, PS− items from the FactBank training set plus some randomly chosen Uu, CT+, and CT− items. (We did not take any CTu sentences into account, as there are not enough of them to support experimentation.) The annotators were asked to decide whether they thought the boldfaced event described in the sentence did (or will) happen. Thus the judgments are from the *reader's perspective*, and not from the *author's or participants' perspective*, as in the original FactBank annotations. We used Saurí's seven-point annotation scheme (removing CTu). To ensure that the workers understood the task, we first gave them four mandatory training items—simple non-corpus examples designed to help them conceptualize the annotation categories properly. The sentences were presented in blocks of 26 items, three of which were "tests" very similar to the training items, included to ensure that the workers were careful. We discarded data from two Turkers because they did not correctly tag the three test sentences.[3]

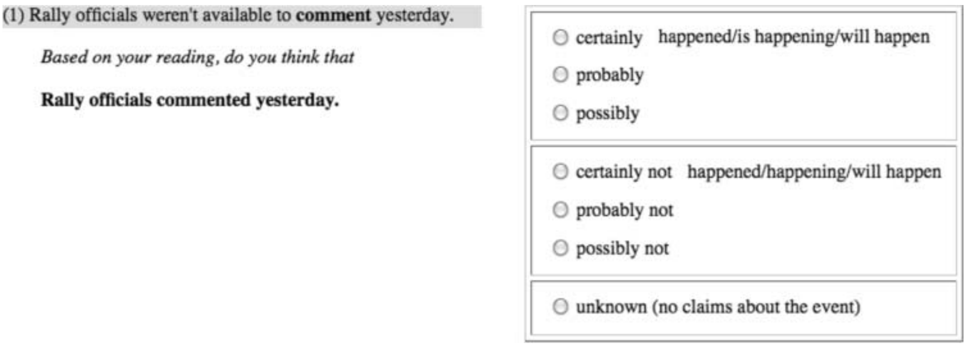
Like Saurí, we did not take the discourse structure into account: Turkers saw only disconnected sentences and judged the event sentence by sentence. Subsequent mentions of the same event in a discourse can, however, lead a listener to revise a veridicality judgment already posed for that event. For instance in Example (4) from Saurí (2008, page 56), a reader's veridicality judgment about the *tipped off* event will probably change when reading the second sentence.

Example 4
Yesterday, the police denied that drug dealers were **tipped off** before the operation. However, it emerged last night that a reporter from London Weekend Television unwittingly **tipped off** residents about the raid when he phoned contacts on the estate to ask if there had been a raid—before it had actually happened.

Here, though, we concentrate on the sentence level, and leave the issue of discourse structure for future work. In other words, we capture the reader's judgment about the veridicality of an event after each sentence, independent on whether the judgment will be revised when later information is read. This is partly to facilitate comparisons with FactBank and partly because we are presently unsure how to computationally model the effects of context in this area.

Figure 1 shows how the items were displayed to the Turkers. We rephrased the event under consideration (the bold sentence in Figure 1), because it is not always straightforward to identify the intended rephrasing. Following Saurí, we refer to this rephrasing process as *normalization*. The normalization strips out any polarity and

[3] The data are available at `http://christopherpotts.net/ling/data/factbank/`.



(1) Rally officials weren't available to **comment** yesterday.

Based on your reading, do you think that

Rally officials commented yesterday.

○ certainly happened/is happening/will happen
○ probably
○ possibly

○ certainly not happened/happening/will happen
○ probably not
○ possibly not

○ unknown (no claims about the event)

Figure 1
Design of the Mechanical Turk experiment.

modality markers to focus only on the core event talked about. For example, in *Police **gave** no details*, we needed to make sure that workers evaluated the positive form ("Police gave details"), rather than the negative one ("Police gave no details"). Similarly, in *Hudson's Bay Co. announced terms of a previously proposed rights issue that is expected to **raise** about 396 million Canadian dollars (US$337 million) net of expenses*, the normalization will remove the modality marker *is expected* ("the proposed rights issue will raise about 396 million Canadian dollars net of expenses"). We followed Saurí's extensive guidelines for this rephrasing process (Saurí 2008, pages 218–222).

We collected 10 annotations for each of the 642 events. A total of 177 Turkers participated in the annotations. Most Turkers did just one batch of 23 non-test examples; the mean number of annotations per Turker was 44, and they each annotated between 23 and 552 sentences. Table 2 reports Fleiss kappa scores (Fleiss 1971) using the full seven-category scheme. These scores are conservative because they do not take into account the fact that the scale is partially ordered, with CT+, PR+, and PS+ forming a "positive" category, CT−, PR−, and PS− forming a "negative" category, and Uu remaining alone. The overall Fleiss kappa for this three-category version is much higher (0.66), reflecting the fact that many of the disagreements were about degree of confidence (e.g., CT+ vs. PR+) rather than the basic veridicality judgment of "positive", "negative", or "unknown". At least 6 out of 10 Turkers agreed on the same tag for 500 of the

Table 2
Fleiss kappa scores with associated p-values.

	κ	p **value**
CT+	0.63	< 0.001
CT−	0.80	< 0.001
PR+	0.41	< 0.001
PR−	0.34	< 0.001
PS+	0.40	< 0.001
PS−	0.12	< 0.001
Uu	0.25	< 0.001
Overall	0.53	< 0.001

642 sentences (78%). For 53% of the examples, at least 8 Turkers agreed with each other, and total agreement is obtained for 26% of the data (165 sentences).

2.3 An Alternative Scale

One of our goals is to assess whether FactBank's seven-category scheme is the right one for the task. To this end, we also evaluated whether a five-tag version would increase agreement and perhaps provide a better match with readers' intuitions. Logically, PR− is equivalent to PS+, and PS− to PR+, so it seemed natural to try to collapse them into a two-way division between "probable" and "possible". We thus ran the MTurk experiment again with the five-point scheme in Table 3.

The five-point scheme led to lower agreement between Turkers. Globally, the PR− items were generally mapped to "no", and PS− to either "no" or "unknown". Some Turkers chose the expected mappings (PS− to "probable" and PR− to "possible"), but only very rarely. This is explicable in terms of the pragmatics of veridicality judgments. Though PR− may be logically equivalent to PS+, and PS− to PR+, there are important pragmatic differences between giving a positive judgment and giving a negative one. For example, in Example (5), speaker B will not infer that he can possibly get a further discount, even if "Probably not" is consistent with "Possibly". Conversely, had the answer been "Possibly", A would have remained hopeful.

Example 5

A: Is it possible to get further discount on the rate?

B: Probably not.

In sum, there seems to be a very intuitive notion of veridicality along the partially ordered scale proposed by Saurí. In their work on assessing the degree of event certainty to which an author commits, Rubin, Liddy, and Kando (2005) used the following five-point scale: *absolute*, *high*, *moderate*, *low*, and *uncertain*. They did not obtain very high inter-annotator agreement ($\kappa = 0.41$). Saurí hypothesized that their low agreement is due to a fuzzy approach and the lack of precise guidelines. Rubin, Liddy, and Kando had no clear identification of certainty markers, and no explicit test for distinguishing different degrees of certainty (unlike Saurí). In our experiment, however, the guidelines were similarly loose: Turkers were instructed only to "read 30 sentences and decide whether the events described in these sentences did (or will) happen." They were not asked to limit their attention to just the information in the sentence, and they were not

Table 3
An alternative five-tag annotation scheme.

Category	Original
yes	CT+
probable	PR+/PS−
possible	PS+/PR−
no	CT−
unknown	Uu

given any mappings between linguistic markers and veridicality values. Nonetheless, Turkers reached good agreement levels in assessing event veridicality on a seven-point scale. We conclude from this that Saurí's scale comes closer than its competitors to capturing reader intuitions about veridicality. The good agreement mirrors the general high inter-annotator agreement levels that have been found for the Recognizing Textual Entailment task (Manning 2006), perhaps reflecting the fact that judging inference and veridicality in context is a natural, everyday human task.

Diab et al. (2009) annotated a 10,000-word corpus for what they call "committed beliefs": whether the author of the sentence indicates with linguistic means that he believes or disbelieves that the event described by the sentence is a fact. Thus, in essence, the annotations assess the degree of event certainty to which an author commits, as in Rubin's work. They use a three-point scale: *committed belief*, *non-committed belief*, and *not applicable*. An example of *committed belief* is *GM has laid off workers*. Affirmative sentences in the future are also considered as *committed belief* (e.g., *GM will lay off workers*). Sentences with modals and events embedded under reported verbs are annotated as *non-committed belief*. The third category, *not applicable*, consists of sentences expressing desire (*Some wish GM would lay of workers*), questions (*Many wonder if GM will lay off workers*), and requirements (*Lay off workers!*). The corpus covers different genres (newswire, e-mail, blog, dialogue). The inter-annotator agreement was high (95%). Prabhakaran, Rambow, and Diab (2010) used the corpus to automatically tag committed beliefs according to that three-point scale. This, too, is an important resource, but it is difficult to compare it with our own task, for two reasons. First, the annotators sought to prevent world knowledge from influencing their annotations, which is concerned only with linguistic markers. Second, the category *non-committed belief* conflates the possible, probable, and unknown categories of our corpus (Saurí's). Though some work in the biomedical domain (i.e., Hobby et al. 2000) suggests that the distinction between possible and probable is hard to make, we did not want to avoid it, because people routinely make such fine-grained modal distinctions when assessing claims. What's more, the approach we develop allows us to quantify the degree to which such judgments are in fact variable and uncertain.

3. Lessons from the New Annotations

This section presents two kinds of high-level analysis of our annotations. We first compare them with FactBank annotations for veridicality according to the author, identifying places where the annotations point to sharp divergences between sentence meaning and utterance meaning. We then study the full distribution of annotations we received (10 per sentence), using them to highlight the uncertainty of veridicality judgments. Both of these discussions deeply inform the modeling work of Section 4.

3.1 The Impact of Pragmatic Enrichment

Although the MTurk annotations largely agree with those of FactBank, there are systematic differences between the two that are indicative of the ways in which pragmatic enrichment plays a role in assessing veridicality. The goal of this section is to uncover those differences. To sharpen the picture, we limit attention to the sentences for which there is a majority-vote category, that is, at least 6 out of 10 Turkers agreed on the annotation. This threshold was met for 500 of the 642 examples.

Table 4
Inter-annotator agreement comparing FactBank annotations with MTurk annotations. The data are limited to the 500 examples in which at least 6 of the 10 Turkers agreed on the label, which is then taken to be the true MTurk label. The very poor value for PS− derives from the fact, in this subset, that label was chosen only once in FactBank and not at all by our annotators.

	κ	p **value**
CT+	0.37	< 0.001
PR+	0.79	< 0.001
PS+	0.86	< 0.001
CT−	0.91	< 0.001
PR−	0.77	< 0.001
PS−	−0.001	$= 0.982$
Uu	0.06	$= 0.203$
Overall	0.60	< 0.001

Table 4 uses kappa scores to measure the agreement between FactBank and our annotations on this 500-sentence subset of the data. We treat FactBank as one annotator and our collective Turkers as a second annotator, with the majority label the correct one for that annotator. What we see is modest to very high agreement for all the categories except Uu. The agreement level is also relatively low for CT+. The corresponding confusion matrix in Table 5 helps explicate these numbers. The Uu category is used much more often in FactBank than by Turkers, and the dominant alternative choice for the Turkers was CT+. Thus, the low score for Uu also effectively drops the score for CT+. The question is why this contrast exists. In other words, why do Turkers choose CT+ where FactBank says Uu?

The divergence can be traced to the way in which lexicalist theories handle events embedded under attitude predicates like *say*, *report*, and *indicate*: any such embedded event is tagged Uu in FactBank. In our annotations, readers are not viewing the veridicality of reported events as unknown. Instead they are sensitive to a wide range of syntactic and contextual features, including markers in the embedded clause, expectations about the subject as a source for the information conveyed by the embedded clause, and lexical competition between the author's choice of attitude predicate and its alternatives. For example, even though the events in Example (6) are all embedded

Table 5
Confusion matrix comparing the FactBank annotations (rows) with our annotations (columns).

					MTurk				
		CT+	PR+	PS+	CT−	PR−	PS−	Uu	Total
FactBank	CT+	54	2	0	0	0	0	0	56
	PR+	4	63	2	0	0	0	0	69
	PS+	1	1	55	0	0	0	2	59
	CT−	5	0	0	146	0	0	2	153
	PR−	0	0	0	0	5	0	1	6
	PS−	0	0	0	0	0	0	1	1
	Uu	94	18	9	12	2	0	21	156
	Total	158	84	66	158	7	0	27	500

under an attitude predicate (*say*), the events in Examples (6a) and (6b) are assessed as certain (CT+), whereas the words *highly confident* in Example (6c) trigger PR+, and *may* in Example (6d) leads to PS+.

Example 6

(a) Magna International Inc.'s chief financial officer, James McAlpine, **resigned** and its chairman, Frank Stronach, is stepping in to help turn the automotive-parts manufacturer around, the company said.

Normalization: James McAlpine resigned

Annotations: CT+: 10

(b) In the air, U.S. Air Force fliers say they have **engaged** in "a little cat and mouse" with Iraqi warplanes.

Normalization: U.S. Air Force fliers have engaged in "a little cat and mouse" with Iraqi warplanes

Annotations: CT+: 9, PS+: 1

(c) Merieux officials said last week that they are "highly confident" the offer will be **approved**.

Normalization: the offer will be approved

Annotations: PR+: 10

(d) U.S. commanders said 5,500 Iraqi prisoners were taken in the first hours of the ground war, though some military officials later said the total may have **climbed** above 8,000.

Normalization: the total Iraqi prisoners climbed above 8,000

Annotations: PS+: 7, PR+: 3

In Example (6a), *the company said* is a parenthetical modifying the main clause (Ross 1973). Asher (2000), Rooryck (2001), and Simons (2007) argue that such constructions often mark the evidential source for the main-clause information, rather than embedding it semantically. In terms of our annotations, this predicts that CT+ or CT− judgments will be common for such constructions, because they become more like two-part meanings: the main-clause and the evidential commentary.

To test this hypothesis, we took from the Turkers' annotations the subset of sentences tagged Uu in FactBank where the event is directly embedded under an attitude verb or introduced by a parenthetical. We also removed examples where a modal auxiliary modified the event, because those are a prominent source of non-CT annotations independently of attitude predication. This yielded a total of 78 sentences. Of these, 33 are parenthetical, and 31 (94%) of those are tagged CT+ or CT−. Of the remaining

45 non-parenthetical examples, 42 (93%) of those are tagged CT+ or CT−. Thus, both parenthetical and non-parenthetical verbs are about equally likely to lead to a CT tag.[4]

This finding is consistent with the evidential analysis of such parentheticals, but it suggests that standard embedding can function pragmatically as an evidential as well. This result is expected under the analysis of Simons (2007) (see also Frazier and Clifton 2005; Clifton and Frazier 2010). It is also anticipated by Karttunen (1973), who focuses on the question of whether attitude verbs are *plugs* for presuppositions, that is, whether presuppositions introduced in their complements are interpreted as semantically embedded. He reviews evidence suggesting that these verbs can be veridical with respect to such content, but he tentatively concludes that these are purely pragmatic effects, writing, "we do not seem to have any alternative except to classify all propositional attitude verbs as plugs, although I am still not convinced that this is the right approach" (page 190). The evidence of the present article leads us to agree with Karttunen about this basic lexicalist classification, with the major caveat that the utterance meanings involved are considerably more complex. (For additional discussion of this point, see Zaenen, Karttunen, and Crouch [2005] and Manning [2006].)

There are also similarities between the FactBank annotations and our own in the case of Uu. As in FactBank, antecedents of conditionals (7), generic sentences (8), and clear cases of uncertainty with respect to the future (9) were tagged Uu by a majority of Turkers.

Example 7

(a) If the heavy outflows **continue**, fund managers will face increasing pressure to sell off some of their junk to pay departing investors in the weeks ahead.

Normalization: the heavy outflows will continue

Annotations: Uu: 7, PS+: 2, CT+: 1

(b) A unit of DPC Acquisition Partners said it would seek to liquidate the computer-printer maker "as soon as possible," even if a merger isn't **consummated**.

Normalization: a merger will be consummated

Annotations: Uu: 8, PS+: 2

Example 8

When prices are tumbling, they must be **willing** to buy shares from sellers when no one else will.

Normalization: they are willing to buy shares

Annotations: Uu: 7, PR+: 2, PS+: 1

[4] We do not regard this as evidence that there is no difference between parenthetical and non-parenthetical uses when it comes to veridicality, but rather only that the categorical examples do not reveal one. Indeed, if we consider the full distribution of annotations, then a linear model with Parenthetical and Verb predicting the number of CT tags reveals Parenthetical to be positively correlated with CT+ (coefficient estimate = 1.36, p = 0.028).

Example 9

(a) The program also calls for **coordination** of economic reforms and joint improvement of social programs in the two countries.

Normalization: there will be coordination of economic reforms and joint improvement of social programs in the two countries

Annotations: Uu: 7, PR+: 2, PS+: 1

(b) But weak car sales raise questions about future **demand** from the auto sector.

Normalization: there will be demand from the auto sector

Annotations: Uu: 6, PS+: 2, CT+: 1, PR−: 1

Another difference between FactBank and the Turkers is the more nuanced categories for PS and PR events. In FactBank, markers of possibility or probability, such as *could* or *likely*, uniquely determine the corresponding tag (Saurí 2008, page 233). In contrast, the Turkers allow the bias created by these lexical items to be swayed by other factors. For example, the auxiliary *could* can trigger a possible or an unknown event (10). In FactBank, all such sentences are marked PS+.

Example 10

(a) They aren't being allowed to leave and could **become** hostages.

Normalization: they will become hostages

Annotations: PS+: 10

(b) Iraq could start **hostilities** with Israel either through a direct attack or by attacking Jordan.

Normalization: there will be hostilities

Annotations: Uu: 6, PS+: 3, PR+: 1

Similarly, *expected* and *appeared* are often markers of PR events. Whereas it is uniquely so in FactBank, however, our annotations show much shifting to PS. Examples (11) and (12) highlight the contrast: It seems likely that the annotators simply have different overall expectations about the forecasting described in each example, a high-level pragmatic influence that does not attach to any particular lexical item.

Example 11

(a) Big personal computer makers are developing 486-based machines, which are expected to **reach** the market early next year.

Normalization: 486-based machines will reach the market early next year

Annotations: PR+: 10

(b) Beneath the tepid news-release jargon lies a powerful threat from the brewing giant, which last year accounted for about 41% of all U.S. beer sales and is expected to see that **grow** to 42.5% in the current year.

Normalization: there will be growth to 42.5% in the current year

Annotations: PS+: 6, PR+: 3, CT+: 1

Example 12

(a) Despite the lack of any obvious successors, the Iraqi leader's internal power base appeared to be **narrowing** even before the war began.

Normalization: the Iraqi leader's internal power base was narrowing even before the war began

Annotations: PR+: 7, CT+: 1, PS+: 1, PS−: 1

(b) Saddam appeared to **accept** a border demarcation treaty he had rejected in peace talks following the August 1988 cease-fire of the eight-year war with Iran.

Normalization: Saddam accepted a border demarcation treaty

Annotations: PS+: 6, PR+: 2, CT+: 2

Another difference is that nouns appearing in a negative context were tagged as CT+ by the Turkers but as CT− or PR− in FactBank.

Example 13

(a) However, its equity in the net income of National Steel declined to $6.3 million from $10.9 million as a result of softer demand and lost **orders** following prolonged labor talks and a threatened strike.

Normalization: there were orders

Annotations: CT+: 6, PR+: 1, PR−: 1, PS+: 1, Uu: 1

This seems to trace to uncertainty about what the annotation should be when the event involves a change of state (from orders existing to not existing). Saurí and Pustejovsky (2009, page 260) note that noun events were a frequent source of disagreement between the two annotators because the annotation guidelines did not address at all how to deal with them.

3.2 The Uncertainty of Pragmatic Enrichment

For the purpose of comparing our annotations with those of FactBank, it is useful to single out the Turkers' majority-choice category, as we did here. We have 10 annotations for each event, however, which invites exploration of the full distribution of annotations, to see if the areas of stability and variation can teach us something about the nature of speakers' veridicality judgments. In this section, we undertake such an exploration, arguing that the patterns reveal veridicality judgments to be importantly probabilistic, as one would expect from a truly pragmatic phenomenon.

Figure 2 provides a high-level summary of the reaction distributions that our sentences received. The labels on the *y*-axis characterize types of distribution. For example, 5/5 groups the sentences for which the annotators were evenly split between two categories (e.g., a sentence for which 5 Turkers assigned PR+ and 5 assigned PS+, or a sentence for which 5 Turkers chose PR+ and 5 chose Uu). The largest grouping, 10, pools the examples on which all the annotators were in agreement.

We can safely assume that some of the variation seen in Figure 2 is due to the noisiness of the crowd-sourced annotation process. Some annotators might have been inattentive or confused, or simply lacked the expertise to make these judgments (Snow et al. 2008). For example, the well-represented 1/9 and 1/1/8 groups probably represent examples for which veridicality assessment is straightforward but one or two of the annotators did not do a good job. If all the distributions were this skewed, we might feel secure in treating veridicality as categorical. There are many examples for which it seems implausible to say that the variation is due to noise, however. For example, 5/5 groups include sentences like Examples (14) and (15), for which the judgments depend heavily on one's prior assumptions about the entities and concepts involved.

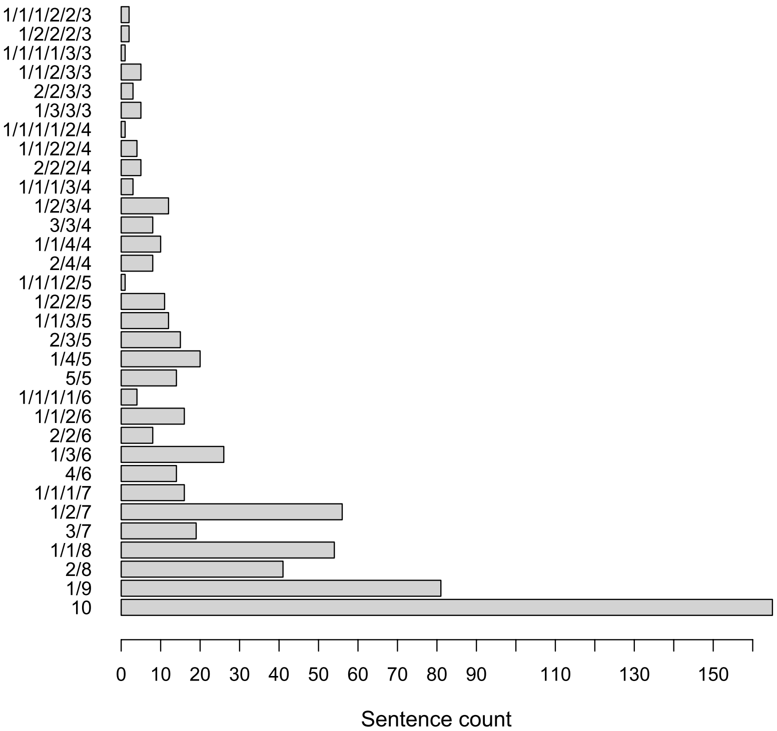


Figure 2
Reaction distributions by type.

Example 14
In a statement, the White House said it would do "whatever is necessary" to ensure **compliance** with the sanctions.

Normalization: there will be compliance with the sanctions

Annotations: Uu: 5, PR+: 5

Example 15
Diplomacy appears to be making headway in **resolving** the United Nations' standoff with Iraq.

Normalization: diplomacy is resolving the United Nations' standoff with Iraq

Annotations: PR+: 5, PS+: 5

The 4/6, 1/4/5, 4/4, and 3/3 groups contain many similarly difficult cases. Combining all of the rows where two categories received at least 3 votes, we get 162 examples, which is 25% of the total data set. Thus, a non-negligible subset of our sentences seem to involve examples where readers' responses are divided, suggesting that there is no unique correct label for them.

Finally, it seems likely that the long tail of very high-entropy distributions at the top of the graph in Figure 2 is owed in large part to the fact that veridicality judgments are often not reachable with confidence, because the utterance is inherently underspecified or because additional contextual information is needed in order to be sure. This, too, suggests to us that it would be foolhardy to assign a unique veridicality label to every example. Of course, situating the sentences in context would reduce some of this uncertainty, but no amount of background information could eliminate it entirely.

When we look more closely at these distributions, we find additional evidence for the idea that veridicality is graded and variable. One of the most striking patterns concerns the question of whether the annotators enriched an example at all, in the following sense. Consider an event that is semantically non-veridical. This could be simply because it is embedded under a non-factive attitude predicate (*say*, *allege*), or an evidential marker (*according to sources*, *it seems*). The semantic strategy for such cases is to pick Uu. Depending on the amount and nature of the contextual information brought to bear on the assessment, however, one might enrich this into one of the positive or negative categories. A cautious positive enrichment would be PS+, for example.

In light of this, it seems promising to look at the subset of 4/6 and 5/5 examples in which one of the chosen categories is Uu, to see what the other choices are like. On the enrichment hypothesis, the other choices should be uniformly positive or negative (up to some noise). Figure 3 summarizes the sentences in our corpus that result in this kind of split. The *y*-axis represents the percentage of non-Uu tags, with the positive values (CT+, PR+, PS+) extending upwards and the negative ones extending downwards. For sentences 1–5 and 18–47 (74% of the total), all of the non-Uu tags were uniform in their basic polarity. What's more, the distributions within the positive and negative portions are highly systematic. In the positive realm, the dominant choice is PS+, the most tentative positive enrichment, followed by PR+, and then CT+. (In the negative realm, CT− is the most represented, but we are unsure whether this supports any definitive conclusions, given the small number of examples.) Our generalization

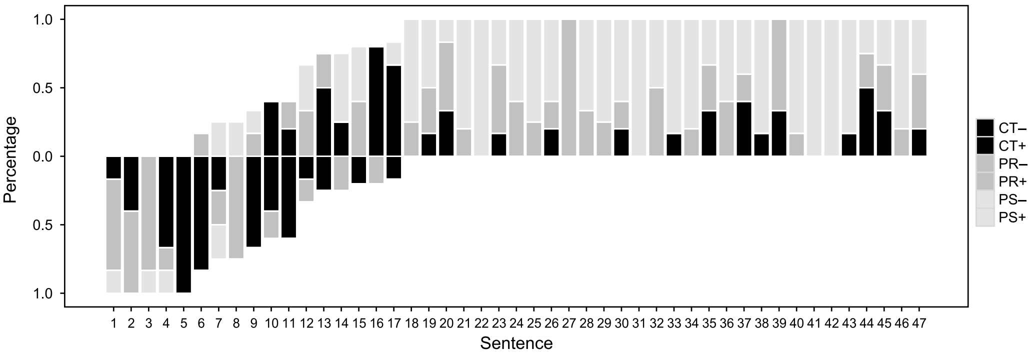


Figure 3
The subset of 4/6 and 5/5 distributions in which one of the dominant categories was Uu. The bars represent the distribution of non-Uu tags for each of these sentences. The top portion depicts the positive tags, and the bottom portion depicts the negative tags.

about these patterns is that enrichment from a semantic Uu baseline is systematic and common, though with interesting variation both in whether it occurs and, if it does, how much.

The full distributions are also informative when it comes to understanding the range of effects that specific lexical items can have on veridicality assessment. To illustrate, we focus on the modal auxiliary verbs *can*, *could*, *may*, *might*, *must*, *will*, *would*.[5] In keeping with lexicalist theories, when they are clausemate to an event, that event is often tagged with one of the PR and PS tags. The relationship is a loose one, however; the modal seems to steer people into these weaker categories but does not determine their final judgment. We illustrate in Example (16) with examples involving *may* in positive contexts.

Example 16

(a) Last Friday's announcement was the first official word that the project was in trouble and that the company's plans for a surge in market share may have been overly **optimistic**.

Normalization: the company's plans have been overly optimistic

Annotations: PS+: 5, PR+: 5

(b) In a letter, prosecutors told Mr. Antar's lawyers that because of the recent Supreme Court rulings, they could expect that any fees collected from Mr. Antar may be **seized**.

Normalization: fees collected from Mr. Antar will be seized

Annotations: PS+: 4, PR+: 6

[5] Other modals, such as *should*, *ought*, and *have to*, are either not well represented in our data or simply absent.

(c) The prospectus didn't include many details about the studio and theme park, although conceptual drawings, released this month, show that it may **feature** several "themed" areas similar to those found at parks built by Walt Disney Co.

Normalization: the park features several "themed" areas similar to those found at parks built by Walt Disney Co.

Annotations: PS+: 4, PR+: 4, CT+:1, Uu:1

Figure 4 summarizes the data for the set of modals on which we could focus. Here, we restrict attention to event descriptions that are clausemate to a modal, effectively taking each modal to be annotated with the distribution of annotations for its clausemate event. We also look only at the positive tags, because the negative ones were too infrequent to provide reliable estimates.

Two types of modals have been recognized in the literature, *weak* and *strong* modals (Wierzbicka 1987; Sæbo 2001; von Fintel and Iatridou 2008; Finlay 2009). Each type has different distribution profiles. As expected, the weak possibility modals *can*, *could*, *may*, and *might* correlate strongly with PS. The other categories are also well-represented for these modals, however, indicating that the contribution of these markers is heavily influenced by other factors. The strong (or necessity) modals *must*, *will*, and *would* are much more evenly distributed across the categories.

The mixed picture for modal auxiliaries seems to be typical of modal markers more generally. We do not have enough data to present a quantitative picture for items like *potentially*, *apparently*, and *partly*, but the following sentences suggest that they are every bit as nuanced in their contributions to veridicality.

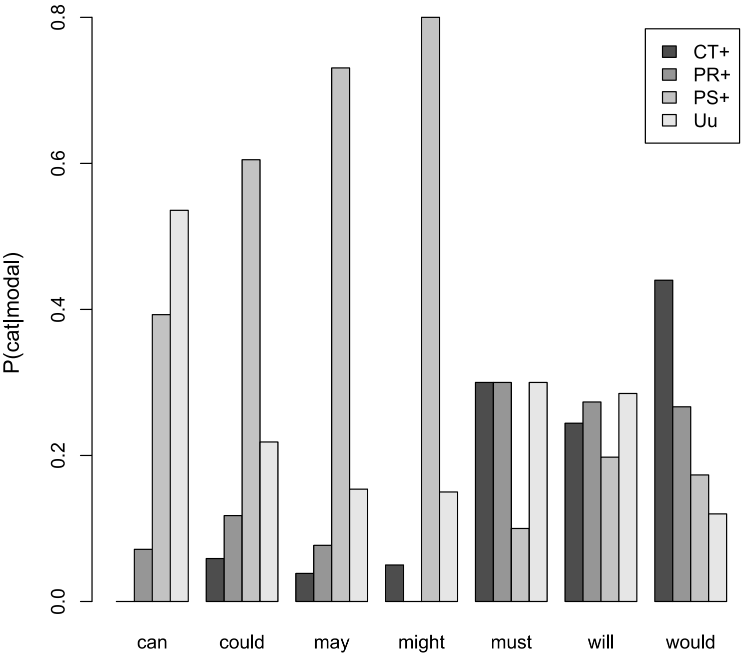


Figure 4
The contribution of modal auxiliaries to veridicality judgments.

Example 17

(a) Anheuser-Busch Cos. said it plans to aggressively discount its major beer brands, setting the stage for a *potentially* bruising price **war** as the maturing industry's growth continues to slow.

Normalization: there will be a bruising price war

Annotations: PS+: 5, PR+: 5

(b) The portfolio unit of the French bank group Credit Lyonnais told stock market regulators that it bought 43,000 shares of Cie. de Navigation Mixte, *apparently* to help **fend** off an unwelcome takeover bid for the company.

Normalization: the 43,000 shares of Cie. de Navigation Mixte will fend off an unwelcome takeover bid for the company

Annotations: PS+: 4, PR+: 4, CT:1, Uu:1

(c) Nonetheless, concern about the chip may have been responsible for a decline of 87.5 cents in Intel's stock to $32 a share yesterday in over-the-counter trading, on volume of 3,609,800 shares, and *partly* **responsible** for a drop in Compaq's stock in New York Stock Exchange composite trading on Wednesday

Normalization: concern about the chip is responsible for a drop in Compaq's stock

Annotations: PS+: 4, PR+: 4, CT+:1, PR−:1

The discussion in this section suggests to us that work on veridicality should embrace variation and uncertainty as part of the characterization of veridicality, rather than trying to approximate the problem as one of basic categorization. We now turn to experiments with a system for veridicality assessment that acknowledges the multi-valued nature of veridicality.

4. A System for Veridicality Assessment

In this section, we describe a maximum entropy classifier (Berger, Della Pietra, and Della Pietra 1996) that we built to automatically assign veridicality. For classification tasks, the dominant tradition within computational linguistics has been to adjudicate differing human judgments and to assign a single class for each item in the training data. In Section 3.2, however, we reviewed the evidence in our annotations that veridicality is not necessarily categorical, by virtue of the uncertainty involved in making pragmatic judgments of this sort. In order to align with our theoretical conception of the problem as probabilistic, we treat each annotator judgment as a training item. Thus, each sentence appears 10 times in our training data.

A maximum entropy model computes the probability of each class *c* given the data *d* as follows:

$$p(c|d,\lambda) = \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}$$

where, for us, the features f_i are indicator functions of a property Φ of the data *d* and a particular class *c*: $f_i(c,d) \equiv \Phi(d) \wedge c = c_k$. The weights λ_i of the features are the parameters of the model chosen to maximize the conditional likelihood of the training data according to the model. The maximum entropy model thus gives us a distribution over the veridicality classes, which will be our output. To assess how good the output of the model is, we will give the log-likelihood of some data according to the model. For comparison, we will also give the log-likelihood for the exact distribution from the Turkers (which thus gives an upper-bound) as well as a log-likelihood for a baseline model which uses only the overall distribution of classes in the training data.

A maximum entropy classifier is an instance of a generalized linear model with a logit link function. It is almost exactly equivalent to the standard multi-class (also called polytomous or multinomial) logistic regression model from statistics, and readers more familiar with this presentation can think of it as such. In all our experiments, we use the Stanford Classifier (Manning and Klein 2003) with a Gaussian prior (also known as L_2 regularization) set to $\mathcal{N}(0,1)$.[6]

4.1 Features

The features were selected through 10-fold cross-validation on the training set.

Predicate classes. Saurí (2008) defines classes of predicates (nouns and verbs) that project the same veridicality value onto the events they introduce. The classes also define the grammatical relations that need to hold between the predicate and the event it introduces, because grammatical contexts matter for veridicality. Different veridicality values will indeed be assigned to *X* in *He doesn't know that X* and in *He doesn't know if X*. The classes have names like ANNOUNCE, CONFIRM, CONJECTURE, and SAY. Like Saurí, we used dependency graphs produced by the Stanford parser (Klein and Manning 2003; de Marneffe, MacCartney, and Manning 2006) to follow the path from the target event to the root of the sentence. If a predicate in the path was contained in one of the classes and the grammatical relation matched, we added both the lemma of the predicate as a feature and a feature marking the predicate class.

[6] The maximum entropy formulation differs from the standard multi-class logistic regression model by having a parameter value for each class giving logit terms for how a feature's value affects the outcome probability relative to a zero feature, whereas in the standard multi-class logistic regression model there are no parameters for one distinguished reference class, and the parameters for other classes say how the value of a feature affects the outcome probability differentially from the reference class. Without regularization, the maximum entropy formulation is overparameterized, and the parameters are unidentifiable; in a regularized setting, however, this is no longer a problem and the maximum entropy formulation then has the advantage that all classes are treated symmetrically, with a simpler symmetric form of model regularization.

World knowledge. For each verb found in the path and contained in the predicate classes, we also added the lemma of its subject, and whether or not the verb was negated. Our rationale for including the subject is that, as we saw in Section 3, readers' interpretations differ for sentences such as *The FBI said it* ***received*** *...* and *Bush said he* ***received*** *...*, presumably because of world knowledge they bring to bear on the judgment. To approximate such world knowledge, we also obtained subject–verb bigram and subject counts from the *New York Times* portion of GigaWord and then included log(subject–verb-counts/subject-counts) as a feature. The intuition here is that some embedded clauses carry the main point of the sentence (Frazier and Clifton 2005; Simons 2007; Clifton and Frazier 2010), with the overall frequency of the elements introducing the embedded clause contributing to readers' veridicality assessments.

General features. We used the lemma of the event, the lemma of the root of the sentence, the incoming grammatical relation to the event, and a general class feature.

Modality features. We used Saurí's list of modal words as features. We distinguished between modality markers found as direct governors or children of the event under consideration, and modal words found elsewhere in the context of the sentence. Figure 4 provides some indication of how these will relate to our annotations.

Negation. A negation feature captures the presence of linguistic markers of negative contexts. Events are considered negated if they have a negation dependency in the graph or an explicit linguistic marker of negation as dependent (e.g., simple negation (*not*), downward-monotone quantifiers (*no*, *any*), or restricting prepositions). Events are also considered negated if embedded in a negative context (e.g., *fail*, *cancel*).

Conditional. Antecedents of conditionals and words clearly marking uncertainty are reliable indicators of the Uu category. We therefore checked for events in an *if*-clause or embedded under markers such as *call for*.

Quotation. Another reliable indicator of the Uu category is quotation. We generated a quotation feature if the sentence opened and ended with quotation marks, or if the root subject was *we*.

In summary, our feature set allows combinations of evidence from various sources to determine veridicality. To be sure, lexical features are important, but they must be allowed to interact with pragmatic ones. In addition, the model does not presume that individual lexical items will contribute in only one way to veridicality judgments. Rather, their contributions are affected by the rest of the feature set.

4.2 Test Data

As test set, we used 130 sentences from the test items in FactBank. We took all the sentences with events annotated PR+ and PS+ at the author level (there are very few), and we randomly chose sentences for the other values (CT+, CT−, and Uu, because the FactBank test set does not contain any PR− and PS− items). Three colleagues provided the normalizations of the sentences following Saurí's guidelines, and the data were then annotated using Mechanical Turk, as described in Section 2. For 112 of the 130 sentences, at least six Turkers agreed on the same value.

5. Results

Table 6 gives log-likelihood values of the classifier for the training and test sets, along with the upper and lower bounds. The upper bound is the log-likelihood of the model that uses the exact distribution from the Turkers. The lower bound is the log-likelihood of a model that uses only the overall rate of each class in our annotations for the training data.

Kullback-Leibler (KL) divergence provides a related way to assess the effectiveness of the classifier. The KL divergence between two distributions is an asymmetric measure of the difference between them. We use Example (6d) to illustrate. For that sentence, the classifier assigns a probability of 0.64 to PS+ and 0.28 to PR+, with very low probabilities for the remaining categories. It thus closely models the gold distribution (PS+: 7/10, PR+: 3/10). The KL divergence is correspondingly low: 0.13. The KL divergence for a classifier that assigned 0.94 probability to the most frequent category (i.e., CT+) and 0.01 to the remaining categories would be much higher: 5.76.

The mean KL divergence of our model is 0.95 (SD 1.13) for the training data and 0.81 (SD 0.91) for the test data. The mean KL divergence for the baseline model is 1.58 (SD 0.57) for the training data and 1.55 (SD 0.47) for the test data. To assess whether our classifier is a statistically significant improvement over the baseline, we use a paired two-sided t-test over the KL divergence values for the two models. The t-test requires that both vectors of values in the comparison have normal distributions. This is not true of the raw KL values, which have approximately gamma distributions, but it is basically true of the log of the KL values: For the model's KL divergences, the normality assumption is very good, whereas for the baseline model there is some positive skew. Nonetheless, the t-test provides a fair way to contextualize and compare the KL values of the two models. By this test, our model improves significantly over the lower bound (two-sided $t = -11.1983$, df = 129, p-value $< 2.2e{-16}$).

We can also compute precision and recall for the subsets of the data where there is a majority vote, that is, where six out of ten annotators agreed on the same label. This allows us to give results per veridicality tag. We take as the true veridicality value the one on which the annotators agreed. The value assigned by the classifier is the one with the highest probability. Table 7 reports precision, recall, and F1 scores on the training and test sets, along with the number of instances in each category. None of the items in our test data were tagged with PR− or PS− and these categories were very infrequent in the training data, so we leave them out. The table also gives baseline results: We

Table 6
Log-likelihood values for the training and test data.

	Train	Test
lower-bound	−10813.97	−1987.86
classifier	−8021.85	−1324.41
upper-bound	−3776.30	−590.75

Table 7
Precision, recall, and F1 on the subsets of the training data (10-fold cross-validation) and test data where there is majority vote, as well as F1 for the baseline.

	Train					Test				
	#	**P**	**R**	**F1**	**Baseline F1**	**#**	**P**	**R**	**F1**	**Baseline F1**
CT+	158	74.3	84.2	78.9	32.6	61	86.9	86.9	86.9	31.8
CT−	158	89.4	91.1	90.2	34.1	31	96.6	90.3	93.3	29.4
PR+	84	74.4	69.1	71.6	19.8	7	50.0	57.1	53.3	6.9
PS+	66	75.4	69.7	72.4	16.7	7	62.5	71.4	66.7	0.0
Uu	27	57.1	44.4	50.0	10.7	6	50.0	50.0	50.0	0.0
Macro-avg		74.1	71.7	72.6	22.8		69.2	71.1	70.0	13.6
Micro-avg		78.6	78.6	78.6	27.0		83.0	83.0	83.0	22.3

used a weighted random guesser, as for the lower-bound given in Table 6. Our results significantly exceed the baseline (McNemar's test, $p < 0.001$).[7]

The classifier weights give insights about the interpretation of lexical markers. Some markers behave as linguistic theories predict. For example, *believe* is often a marker of probability whereas *could* and *may* are more likely to indicate possibility. But as seen in Examples (10) and (16), world knowledge and other linguistic factors shape the veridicality of these items. The greatest departure from theoretical predictions occurs with the SAY category, which is logically non-veridical but correlates highly with certainty (CT+) in our corpus.[8] Conversely, the class KNOW, which includes *know*, *acknowledge*, and *learn*, is traditionally analyzed as veridical (CT+), but in our data is sometimes a marker of possibility, as we discuss in the Conclusion. Our model thus shows that to account for how readers interpret sentences, the space of veridicality should be cut up differently than the lexicalist theories propose.

6. Error Analysis

We focus on two kinds of errors. First, where there is a majority label (a label six or more of the annotators agreed on) in the annotations, we can compare that label with the one assigned the highest probability according to our model. Second, we can study cases where the the annotation distribution diverges considerably from our model's distribution (i.e., cases with a very high KL divergence).

For the majority-label cases, errors of polarity are extremely rare; the classifier wrongly assesses the polarity of only four events, shown in Example (18). Most of the errors are thus in the degree of confidence (e.g., CT+ vs. PR+). The graphs next

[7] McNemar's test assesses whether the proportions of right and wrong predictions for two systems are significantly different. We calculated the test exactly via the binomial distribution. We chose a random guess baseline rather than a choose-most-frequent-class baseline hoping to illustrate a sensible baseline F1 performance for each class, but actually the baseline F1 remains zero for two classes on the test set (there are few items in each class and the random guesser never guessed correctly). In sum, the performance differences are significant for both the training and test sets compared to either of these baselines.

[8] This might be due to the nature of our corpus, namely, newswire, where in the vast majority of the cases, reports are considered true. The situation could be totally different in another genre (such as blogs, for instance).

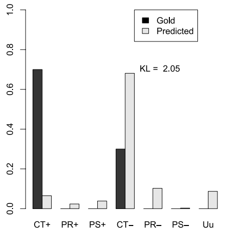
to the examples compare the gold annotation from the Turkers (the black bars) with the distribution proposed by the classifier (the gray bars). The KL divergence value is included to help convey how such values relate to these distributions.

Example 18

(a) Addressing a NATO flag-lowering ceremony at the Dutch embassy, Orban said the occasion indicated the end of the embassy's **mission** of liaison between Hungary and NATO.

Normalization: there is an embassy's mission of liaison between Hungary and NATO

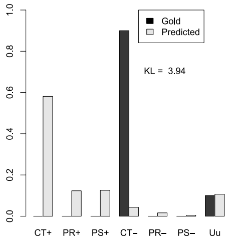
Annotations: CT+:7, CT−: 3



(b) But never before has NATO **reached** out to its former Eastern-bloc enemies.

Normalization: NATO has reached out to its former Eastern-bloc enemies in the past

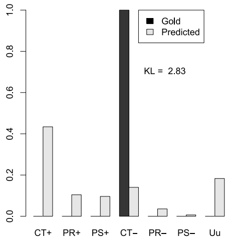
Annotations: CT−: 9, Uu: 1



(c) Horsley **was** not a defendant in the suit, in which the Portland, Ore., jury ruled that such sites constitute threats to abortion providers.

Normalization: Horsley was a defendant in the suit

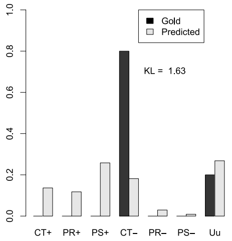
Annotations: CT−: 10



(d) A total of $650,000, meanwhile, is being offered for information leading to the **arrest** of Kopp, who is charged with gunning down Dr. Barnett Slepian last fall in his home in Buffalo.

Normalization: Kopp has been arrested

Annotations: CT−: 8, Uu: 2



When the system missed CT− events, it failed to find an explicit negative marker, as in Example (18b), where (due to a parse error) *never* is treated as a dependent of the verb *have* and not of the *reaching out* event. Similarly, the system could not capture instances in which the negation was merely implicit, as in Example (18d), where the nonveridicality of the arresting event requires deeper interpretation that our feature-set can manage.

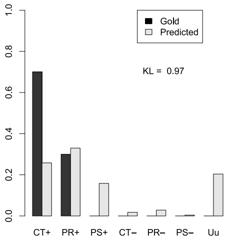
In Example (19), we give examples of CT+ events that are incorrectly tagged PR+, PS+, or Uu by the system because of the presence of a weak modal auxiliary or a verb that lowers certainty, such as *believe*. As we saw in Section 3.2, these markers correlate strongly with the PS categories.

Example 19

(a) The NATO summit, she said, would produce an **initiative** that "responds to the grave threat posed by weapons of mass destruction and their means of delivery."

Normalization: there will be an initiative

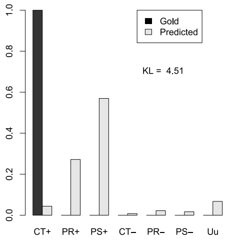
Annotations: CT+: 7, PR+: 3



(b) Kopp, meanwhile, may have approached the border with Mexico, but it is **unknown** whether he crossed into that country, said Freeh.

Normalization: it is unknown whether Kopp crossed into Mexico

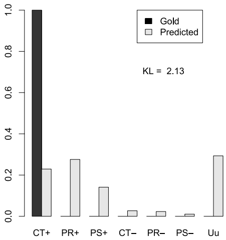
Annotations: CT+: 10



(c) They believe Kopp was driven to Mexico by a female friend after the **shooting**, and have a trail of her credit card receipts leading to Mexico, the federal officials have said.

Normalization: there was a shooting

Annotations: CT+: 10



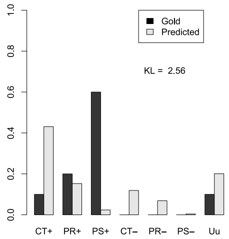
In the case of PR+ and PS+ events, all the erroneous values assigned by the system are CT+. Some explicit modality markers were not seen in the training data, such as *potential* in Example (20a), and thus the classifier assigned them no weight. In other cases, such as Example (20b), the system did not capture the modality implicit in the conditional.

Example 20

(a) Albright also used her speech to articulate a forward-looking vision for NATO, and to defend NATO's potential **involvement** in Kosovo.

Normalization: NATO will be involved in Kosovo

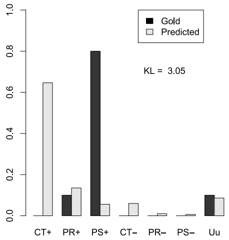
Annotations: PS+: 6, PR+: 2, CT+: 1, Uu: 1



(b) "And we must be resolute in spelling out the consequences of intransigence," she added, referring to the threat of NATO air **strikes** against Milosevic if he does not agree to the deployment.

Normalization: there will be NATO air strikes

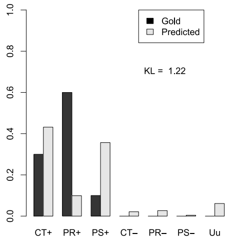
Annotations: PS+: 8, PR+: 1, Uu: 1



(c) But the decision by District Attorney Frank C. Clark to begin presenting evidence to a state grand jury suggests that he has **amassed** enough material to support a criminal indictment for homicide.

Normalization: District Attorney Frank C. Clark has amassed material to support a criminal indictment for homicide

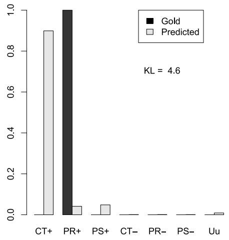
Annotations: PR+: 6, CT+: 3, PS+: 1



(d) The first round of DNA tests on the hair at the FBI Laboratory here established a high probability it **came** from the same person as a hair found in a New Jersey home where James C. Kopp, a 44-year-old anti-abortion protester, lived last year, the official said.

Normalization: the hair came from the same person

Annotations: PR+: 10



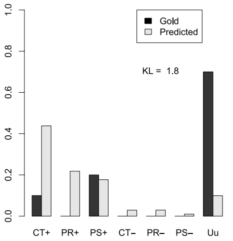
The only Uu events that the system correctly retrieved were antecedents of a conditional. For the other Uu events in Example (21), the system assigned CT+ or PR+. The majority of Uu events proved to be very difficult to detect automatically since complex pragmatic factors are at work, many of them only very indirectly reflected in the texts.

Example 21

(a) Kopp's stepmother, who married Kopp's father when Kopp was in his 30s, said Thursday from her home in Irving, Texas: "I would like to see him come forward and clear his name if he's not guilty, and if he's guilty, to contact a priest and make his amends with society, face what he **did**."

Normalization: Kopp did something

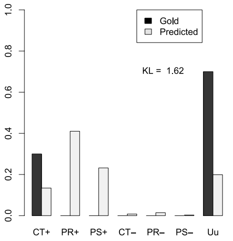
Annotations: Uu: 7, PS+: 2, CT+: 1



(b) Indeed, one particularly virulent anti-abortion Web site lists the names of doctors it says perform abortions, or "crimes against humanity," with a code indicating whether they are "**working**," "wounded" or a "fatality."

Normalization: doctors are working

Annotations: Uu:7, CT+: 3

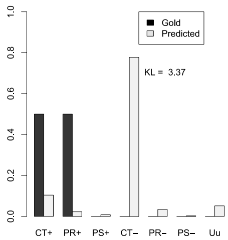


It is also instructive to look at the examples for which there is a large KL divergence between our model's predicted distribution and the annotation distribution. Very often, this is simply the result of a divergence between the predicted and actual majority label, as discussed earlier. Instances like Example (22) are more interesting in this regard, however: These are cases where there was no majority label, as in Example (22a), or where the model guessed the correct majority label but failed to capture other aspects of the distribution, as in Examples (22b) and (22c).

Example 22

(a) On Tuesday, the National Abortion and Reproductive Rights Action League plans to hold a news **conference** to screen a television advertisement made last week, before Slepian died, featuring Emily

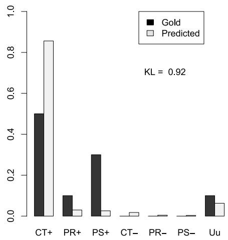
Lyons, a nurse who was badly wounded earlier this year in the bombing of an abortion clinic in Alabama.



Normalization: there will be a news conference to screen a television advertisement

Annotations: CT+: 5, PR+: 5

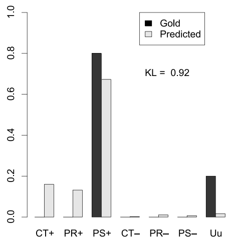
(b) Vacco's campaign manager, Matt Behrmann, said in a statement that Spitzer had "sunk to a new and despicable low by **attempting** to capitalize on the murder of a physician in order to garner votes."



Normalization: Spitzer had attempted to capitalize on the murder of a physician in order to garner votes

Annotations: CT+: 5, PR+: 1, PS+: 3, Uu: 1

(c) Since there is no federal homicide statute as such, the federal officials said Kopp could be **charged** under the recent Freedom of Access to Clinic Entrances Act, which provides for a sentence of up to life imprisonment for someone convicted of physical assaults or threats against abortion providers.



Normalization: Kopp will be charged under the recent Freedom of Access to Clinic Entrances Act

Annotations: PS+: 8, Uu: 2

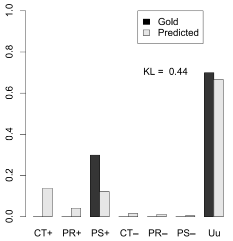
In Example (22a), the classifier is confused by an ambiguity: it treats *hold* as a kind of negation, which leads the system to assign a 0.78 probability to CT−. In Example (22b), there are no features indicating possibility, but a number of SAY-related features are present, which leads to a very strong bias for CT+ (0.86) and a corresponding failure to model the rest of the distribution properly. In Example (23c), the classifier correctly assigns most probability to PS+, but the rest of the probability mass is distributed between CT+ and PR+. This is another manifestation of the problem, noted earlier, that we have very few strong indicators of Uu. The exception to that is conditional antecedents. As a result, we do well with cases like Example (23a), where the event is in a conditional; the classifier assigns 70% of the probability to Uu and 0.15 to PS+.

Example 23

(a) On Monday, Spitzer called for Vacco to revive that unit immediately, vowing that he would do so on his first day in office if **elected**.

Normalization: Spitzer will be elected

Annotations: Uu: 7, PS+: 3



Overall the system assigns incorrect veridicality distributions in part because it misses explicit linguistic markers of veridicality, but also because contextual and pragmatic factors cannot be captured. This is instructive, though, and serves to further support our central thesis that veridicality judgments are not purely lexical, but rather involve complex pragmatic reasoning.

7. Conclusion

Our central goal for this article was to explore veridicality judgments at the level of utterance meaning. To do this, we extended FactBank (Saurí and Pustejovsky 2009) with veridicality annotations that are informed by context and world knowledge (Section 2). Although the two sets of annotations are similar in many ways, their differences highlight areas in which pragmatic factors play a leading role in shaping readers' judgments (Section 3.1). In addition, because each one of our sentences was judged by 10 annotators, we actually have annotation *distributions* for our sentences, which allows us to identify areas of uncertainty in veridicality assessment (Section 3.2). This uncertainty is so pervasive that the problem itself seems better modeled as one of predicting a distribution over veridicality categories, rather than trying to predict a single label. The predictive model we developed (Section 4) is true to this intuition, because it trains on and predicts distributions. All the features of the model, even the basic lexical ones, show the influence of interacting pragmatic factors (Section 5). Although automatically assigning veridicality judgments that correspond to readers' intuitions when pragmatic factors are allowed to play a role is challenging, our classifier shows that it can be done effectively using a relatively simple feature set, and we expect performance to improve as we find ways to model richer contextual features.

These findings resonate with the notion of entailment used in the Recognizing Textual Entailment challenges (Dagan, Glickman, and Magnini 2006), where the goal is to determine, for each pair of sentences $\langle T, H\rangle$, whether T (the *text*) justifies H (the *hypothesis*). The original task definition draws on "common-sense" understanding of language (Chapman 2005), and focuses on how people interpret utterances naturalistically. Thus, these entailments are not calculated over just the information contained in the sentence pairs, as a more classical logical approach would have it, but rather over the full utterance meaning. As a result, they are imbued with all the uncertainty of utterance meanings (Zaenen, Karttunen, and Crouch 2005; Crouch, Karttunen, and Zaenen 2006; Manning 2006). This is strongly reminiscent of our distinction between semantic and pragmatic veridicality. For example, as a purely semantic fact, *might*(S) is non-veridical with regard to *S*. Depending on the nature of *S*, however, the nature of the source, the context, and countless other factors, one might nonetheless infer *S*. This is one of the central lessons of our new annotations.

In an important sense, we have been conservative in bringing semantics and pragmatics together, because we do not challenge the basic veridicality categorizations that come from linguistic and logical work on this topic. Rather, we just showed that those semantic judgments are often enriched pragmatically—for example, from uncertainty to one of the positive or negative categories, or from PS to PR or even CT. The interaction between lexical markers and pragmatic context is also crucial in the case of absolutism: Too many lexical markers conveying certainty might in some cases undermine the speaker's credibility (e.g., in a car salesman pitch) and actually incite mistrust in the hearer/reader. In such instances, lexical markers only are not good indicators of the veridicality of the event, but the pragmatic context of the utterance needs to be taken into account to fully appreciate the interpretation people assign to it. There is, however, evidence suggesting that we should be even more radically pragmatic (Searle 1978; Travis 1996), by dropping the notion that lexical items can be reliably classified once and for all. For example, lexical theories generally agree that *know* is veridical with respect to its sentential complement, and the vast majority of its uses seem to support that claim. There are exceptions, though, as in Example (24) (see also Beaver 2010; Hazlett 2010):

Example 24

(a) For the first time in history, the U.S. has gone to war with an Arab and Muslim nation, and we know a peaceful solution **was in reach**.

(b) Let me tell you something, when it comes to finishing the fight, Rocky and I have a lot in common. I never quit, I never give up, and I know that we're going to **make it together**.
– Hillary Clinton, 1 September 2008.

(c) "That woman who knew I **had dyslexia** – I never interviewed her."
– George W. Bush (*New York Times*, 16 September 2000. Quoted by Miller [2001].)

All of these examples seem to use *know* to report emphatically held belief, a much weaker sense than a factive lexical semantics would predict. Example (24c) is the most striking of the group, because it seems to be pragmatically non-veridical: The continuation is Bush's evidence that the referent of *that woman* could not possibly be in a position to determine whether he is dyslexic. Such examples further emphasize the importance of a pragmatically informed perspective on veridicality in natural language.

One key component of veridicality judgments that we left out in this study is the text provenance. Our data did not allow us to examine its impact because we did not have enough variation in the provenance. All FactBank sentences are from newspaper and newswire text such as the *Wall Street Journal*, the Associated Press, and the *New York Times*. The trustworthiness of the document provenance can affect veridicality judgments, however: People might have different reactions reading a sentence in the *New York Times* versus in a random blog on the Web. We plan to examine and incorporate the role of text provenance in future work.

Acknowledgments

We thank our three anonymous referees for their insights and advice. Thanks also to Anastasia Giannakidou, Jesse Harris, Ed Hovy, Lauri Karttunen, James Pustejovsky, and Annie Zaenen for valuable comments at various stages of the project. A special thanks to Eric Acton, Alex Djalali, and Scott Grimm for their help with the test data. This research was supported in part by ONR grant no. N00014-10-1-0109, ARO grant no. W911NF-07-1-0216, and by a Stanford Interdisciplinary Graduate Fellowship to the first author.

References

Asher, Nicholas. 2000. Truth conditional discourse semantics for parentheticals. *Journal of Semantics*, 17(1):31–50.

Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Barwise, Jon. 1981. Scenes and other situations. *The Journal of Philosophy*, 78(7):369–397.

Beaver, David. 2010. Have you noticed that your belly button lint colour is related to the colour of your clothing? In Rainer Bäuerle, Uwe Reyle, and Thomas Ede Zimmermann, editors, *Presuppositions and Discourse: Essays Offered to Hans Kamp*. Elsevier, Philadelphia, PA, pages 65–99.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Chapman, Siobhan. 2005. *Paul Grice: Philosopher and Linguist*. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire.

Chapman, Wendy W., Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Clifton, Charles Jr. and Lyn Frazier. 2010. Imperfect ellipsis: Antecedents beyond syntax? *Syntax*, 13(4):279–297.

Crouch, Richard, Lauri Karttunen, and Annie Zaenen. 2006. Circumscribing is not excluding: A reply to Manning. Ms., Palo Alto Research Center, Palo Alto, CA.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alché-Buc, editors, *Machine Learning Challenges, Lecture Notes in Computer Science*, volume 3944. Springer-Verlag, New York, pages 177–190.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, pages 449–454.

Diab, Mona, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, Singapore, pages 68–73.

Elkin, Peter L., Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom, and Dietlind L. Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 5(13).

Farkas, Richárd, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task*, Uppsala, Sweden, pages 1–12.

Finlay, Stephen. 2009. Oughts and ends. *Philosophical Studies*, 143(3):315–340.

Fleiss, Joseph I. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Frazier, Lyn and Charles Clifton, Jr. 2005. The syntax-discourse divide: Processing ellipsis. *Syntax*, 8(1):121–174.

Giannakidou, Anastasia. 1994. The semantic licensing of NPIs and the Modern Greek subjunctive. In *Language and Cognition 4, Yearbook of the Research Group for Theoretical and Experimental Linguistics*. University of Groningen, The Netherlands, pages 55–68.

Giannakidou, Anastasia. 1995. Weak and strong licensing: Evidence from Greek. In Artemis Alexiadou, Geoffrey Horrocks, and Melita Stavrou, editors, *Studies in Greek Syntax*. Kluwer, Dordrecht, pages 113–133.

Giannakidou, Anastasia. 1999. Affective dependencies. *Linguistics and Philosophy*, 22(4):367–421.

Giannakidou, Anastasia. 2001. The meaning of free choice. *Linguistics and Philosophy*, 24(6):659–735.

Hazlett, Allan. 2010. The myth of factive verbs. *Philosophy and Phenomenological Research*, 80(3):497–522.

Hobby, Jonathan L., Brian D. M. Tom, C. Todd, Philip W. P. Bearcroft, and Adrian K. Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, 73(873):999–1001.

Huang, Yang and Henry J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311.

Karttunen, Lauri. 1973. Presuppositions and compound sentences. *Linguistic Inquiry*, 4(2):169–193.

Karttunen, Lauri and Annie Zaenen. 2005. Veridicity. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, Extracting and Reasoning about Time and Events*, number 05151 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on BioNLP: Shared Task*, Boulder, Colorado, pages 1–9.

Kiparsky, Paul and Carol Kiparsky. 1970. Facts. In M. Bierwisch and K. E. Heidolph, editors, *Progress in Linguistics*. Mouton. The Hague, Paris, pages 143–173.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association of Computational Linguistics*, Sapporo, Japan, pages 423–430.

Levinson, Stephen C. 1995. Three levels of meaning: Essays in honor of Sir John Lyons. In Frank R. Palmer, editor, *Grammar and Meaning*. Cambridge University Press, Cambridge, pages 90–115.

Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, MA.

Manning, Christopher D. 2006. Local textual inference: it's hard to circumscribe, but you know it when you see it—and NLP needs it. Ms., Stanford University.

Manning, Christopher D. and Dan Klein. 2003. Optimization, maxent models, and conditional estimation without magic. In *Tutorial at HLT-NAACL 2003 and ACL 2003*. Available at `http://nlp.stanford.edu/software/classifier.shtml`.

Miller, Mark Crispin. 2001. *The Bush Dyslexicon*. W. W. Norton and Company, New York, NY.

Montague, Richard. 1969. On the nature of certain philosophical entities. *The Monist*, volume 2, pages 159–194.

Morante, Roser and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Boulder, Colorado, pages 21–29.

Morante, Roser and Caroline Sporleder, editors. 2010. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, Sweden.

Prabhakaran, Vinodkumar, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Proceedings of COLING 2010: Poster Volume*, Beijing, China, pages 1014–1022.

Pustejovsky, James, Marc Verhagen, Roser Saurí, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. 2006. Timebank 1.2. Linguistic Data Consortium, Philadelphia, PA.

Pyysalo, Sampo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50). doi:10.1186/1471-2105-8-50.

Rieh, Soo Young. 2010. Credibility and cognitive authority of information. In M. Bates and M. N. Maack, editors, *Encyclopedia of Library and Information Sciences, 3rd ed.*, Taylor and Francis Group, LLC, New York, pages 1337–1344.

Rooryck, Johan. 2001. Evidentiality, Part I. *Glot International*, 5(4):3–11.

Ross, John Robert. 1973. Slifting. In Maurice Gross, Morris Halle, and Marcel-Paul Schützenberger, editors, *The Formal Analysis of Natural Languages*. Mouton de Gruyter, The Hague, pages 133–169.

Rubin, Victoria L. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Proceedings of the NAACL-HLT 2007*, Rochester, NY, pages 141–144.

Rubin, Victoria L., Elizabeth D. Liddy, and Noriko Kando. 2005. Certainty identification in texts: Categorization model and manual tagging results. In J. G. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series)*, Springer-Verlag, New York, pages 61–76.

Sæbo, Kjell Johan. 2001. Necessary conditions in a natural language. In Caroline Féry and Wolfgang Sternefeld, editors, *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*. Akademie Verlag, Berlin, pages 427–449.

Saurí, Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Computer Science Department, Brandeis University.

Saurí, Roser and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Searle, John R. 1978. Literal meaning. *Erkenntnis*, 13:207–224.

Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, pages 254–263.

Szarvas, György, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, Columbus, OH, pages 38–45.

Travis, Charles. 1996. Meaning's role in truth. *Mind*, 105(419):451–466.

von Fintel, Kai and Sabine Iatridou. 2008. How to say *ought* in foreign: The composition of weak necessity modals. In Jacqueline Guéron and Jacqueline Lecarme, editors, *Studies in Natural Language and Linguistic Theory*, volume 75. Springer, Berlin, pages 115–141.

Wierzbicka, Anna. 1987. The semantics of modality. *Folia Linguistica*, 21(1):25–43.

Zaenen, Annie, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *ACL Workshop on Empirical Modelling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, MI.

Zwarts, Frans. 1995. Nonveridical contexts. *Linguistic Analysis*, 25(3–4):286–312.