Lang Resources & Evaluation
DOI 10.1007/s10579-009-9089-9

FactBank: a corpus annotated with event factuality

Roser Saurí · James Pustejovsky

Abstract Recent work in computational linguistics points out the need for systems to be sensitive to the veracity or *factuality* of events as mentioned in text; that is, to recognize whether events are presented as corresponding to actual situations in the world, situations that have not happened, or situations of uncertain interpretation. Event factuality is an important aspect of the representation of events in discourse, but the annotation of such information poses a representational challenge, largely because factuality is expressed through the interaction of numerous linguistic markers and constructions. Many of these markers are already encoded in existing corpora, albeit in a somewhat fragmented way. In this article, we present FACTBANK, a corpus annotated with information concerning the factuality of events. Its annotation has been carried out from a descriptive framework of factuality grounded on both theoretical findings and data analysis. FactBank is built on top of TimeBank, adding to it an additional level of semantic information.

Keywords Event factuality · Modality · Certainty · Subjectivity analysis · Corpus creation · TimeBank

1 Introduction

Many recent efforts towards corpus construction have been devoted to encoding a variety of semantic information structures in text. For example, much work has gone toward annotating the basic units that configure propositions (PropBank, FrameNet)

R. Saurí (✉) · J. Pustejovsky
Laboratory for Linguistics and Computation, Computer Science Department,
Brandeis University, Waltham, MA, USA
e-mail: roser@cs.brandeis.edu

J. Pustejovsky
e-mail: jamesp@cs.brandeis.edu

Published online: 07 May 2009

and the relations that hold between them at the discourse level (RST Corpus, Penn Discourse TreeBank, GraphBank), as well as specific knowledge that has proved fundamental in tasks requiring some degree of text understanding, such as temporal information (TimeBank) and opinion expressions (MPQA Opinion Corpus).[1] Research is moving now towards finding common platforms for unifying these in an optimal way, e.g., Pradhan et al. (2007) and Verhagen et al. (2007). It therefore seems we are at a point where the first elements for text understanding can be brought together. Nonetheless, some major semantic levels are still missing from the picture. Current work from different disciplines within NLP point out, for example, the need for systems to be sensitive to information conveying whether events mentioned in text correspond to real situations in the world or, instead, to situations of uncertain status. We refer to this level of information as *event factuality*.

Together with argument structure as well as tense, aspect, and temporal ordering information, the factuality of an event is an important component for interpreting events in discourse: inferences derived from events that have not happened (or that are only possible) are different from those derived from events judged as factual in nature.

The need for this further type of information is demonstrated in several disciplines within NLP. For example, Karttunen and Zaenen (2005) discuss the relevance of veridicity for information extraction (IE), while Saurí et al. (2006a) frame it within the context of tasks requiring some degree of narrative understanding, e.g., question answering (QA). Factuality is critical also in the area of opinion detection (Wiebe et al. 2005), given that the same situation can be presented as a fact in the world, a mere possibility, or a counterfact according to different sources. In the domain of textual entailment, factuality-related information (modality, intensional contexts, etc.) has been taken as a basic feature in some systems participating in (or using the data from) previous PASCAL RTE challenges; e.g., Tatu and Moldovan (2005), de Marneffe et al. (2006a), and Snow and Vanderwende (2006). Most significantly, the system that obtained the best absolute result in the three RTE challenges, scoring an 80% accuracy (Hickl and Bensley 2007), is based on identifying the set of publicly-expressed beliefs of the author; that is, on the author's commitments of how things are in the world according to what is expressed in text. Finally, the factuality of events has been also pointed out as an important information component in highly domain-oriented areas such as bioinformatics (Light et al. 2004).

Despite the important role of this level of information, few systems in the areas of IE, QA, or textual entailment have taken advantage of the structural clues and subsequent semantic interpretations that are possible regarding factuality, largely because identifying the factuality of events presents challenges at several levels.

[1] The main references for these corpora are: PropBank (Palmer et al. 2005), FrameNet (Baker et al. 1998), RST Corpus (Carlson et al. 2003), Penn Discourse TreeBank (Miltsakaki et al. 2004), GraphBank (Wolf and Gibson 2005), TimeBank (Pustejovsky et al. 2006), MPQA Opinion Corpus (Wiebe et al. 2005).

First, event factuality in itself does not constitute a discrete linguistic system. Events in language are couched in terms of a continuum that ranges from truly factual to counterfactual, passing through a whole spectrum of shades of modality that languages accommodate in different ways, depending on the grammatical resources they have available. The challenge for annotation is, therefore, to find an expressive enough set of discrete factuality values that is grounded on linguistic intuitions but also supported by commonsense reasoning. Only in this way can we aim at both obtaining an acceptable degree of interannotation agreement (and thus, of annotation consistency) and, at the same time, establishing a solid model of factuality with potential cross-linguistic validity.

A second main challenge in annotating this type of information is presented by the fact that factuality is expressed through a complex interaction of many different aspects of the overall linguistic expression. It involves polarity (events can be presented as positive or negative) as well as epistemic modality, which expresses the degree of certainty of a source regarding what is asserted. Other information at play is evidentiality (e.g., a *seen* event is presented with a factuality degree stronger than that of an event *reported* by someone else), and mood (e.g., indicative vs. subjunctive). Factuality is also a component in the semantics of specific syntactic structures with presuppositional effects (e.g., appositions, non-restrictive relative clauses), as well as certain types of predicates, most notoriously the so-called factive and implicative predicates, but also others.

Some of these factuality markers are already encoded in some existing corpora, but always in a somewhat incomplete manner, either focusing only on identifying the factuality markers while disregarding the factuality value that results from their interaction, or characterizing the factual nature of the expressed event by means of a fairly shallow model, unable to account for distinctions in a general, predictive way.

In this article, we introduce FACTBANK, a corpus of events annotated with factuality information. The annotation language is grounded on established linguistic analyses of the phenomenon, which facilitated creating a battery of discriminatory tests for distinguishing between factuality values. Consequently, this resulted in a relatively high interannotation agreement ($\kappa_{cohen} = 0.81$). FactBank has been built on top of TimeBank. Together, TimeBank and FactBank offer a double-layered annotation of event factuality: the former encodes most of the basic structural elements expressing factuality information, whereas the latter represents the resulting factuality interpretation.

The article begins by setting the linguistic grounding of the phenomenon: Section 2 defines event factuality as a level of subjectivity information, and identifies the most common linguistic devices to convey it. Then, Sect. 3 presents the main challenges in annotating such information, and Sect. 4 reviews some of the existing corpora annotated with factuality-related information. FactBank is introduced in Sect. 5, where the set of factuality values is presented, together with their discriminatory tests. The corpus is described in some detail and compared to TimeBank. Section 6 addresses the annotation effort while Sect. 7 evaluates and discusses the results.

2 Event factuality, a level of subjective information

2.1 Subjectivity analysis

At a time when work on the basic units of linguistic information (e.g., pos-tagging, morphological analysis, basic syntactic parsing) is reaching a somewhat stable state, the community has begun to gravitate towards semantic and pragmatic-based phenomena. It is under this new focus of interest that, in recent years, an area generally termed *subjectivity* analysis has become active within the field. Beyond data that is inherently linguistic in nature, subjectivity analysis deals with the notion of *perspective*; that is, it concerns discourse participants and their stance with respect to what is conveyed by means of the text. Whenever we use language to talk about situations in the world, we are not only referring to relations among a number of participants which potentially hold at given locations and specific points or periods of time, but we are also expressing our particular stance about them. This constitutes what is called the subjectivity level of the text.

Within linguistics, this level of information has been studied for some time from different traditions. For example, from a more cognitivist perspective, there is considerable work on epistemic (or epistemological) stance in natural language, which is defined as the pragmatic relation between speakers and the things they talk about—i.e., their knowledge. Such relations include aspects like: how that knowledge was acquired, how speakers assess its quality, etc. (Mushin 2001). Epistemic stance can be of different kinds, including: attitude, judgement, or commitment (Biber and Finegan 1989). Within Systemic Functional Linguistics (Halliday 1994), the Appraisal Framework develops a taxonomy of the mechanisms employed for expressing subjective information such as attitude, its polarity, graduation, etc. (Martin and White 2005).

Subjectivity manifests itself along different parameters, and hence it encompasses a diverse set of interrelated research lines within the fields of NLP and data mining. Some work, for instance, is devoted to identifying the author's *affectual (or emotional) state* (e.g., Dave 2003). This area is often referred to as *sentiment analysis*. Another related area focuses on *opinions*. Opinion identification can be performed at different levels of granularity: document (Pang et al. 2002; Turney 2002), clausal and phrasal (e.g., Wiebe et al. 2005; Read et al. 2007; Stoyanov and Cardie 2008), or even lexical (e.g., Wiebe 2000; Riloff et al. 2003; Andreevskaia and Bergler 2006). Furthermore, opinions can be analyzed along different facets, such as the polarity of the attitude being expressed (i.e., positive vs. negative; e.g., Wilson et al. 2005), its rating (i.e., how much positively or negatively the target is evaluated as) (Pang and Lee 2005), and the expression strength (e.g., *neutral, low, medium*, and *high*), which touches on author's affectual state (Wilson et al. 2004).

In addition to affectual and opinion analysis, a further parameter configuring the subjectivity component in discourse deals with *event factuality*. Consider:

(1) a. Jubilant Red Sox fans cheered for players at Fenway Park yesterday.
b. The US Soccer Team may play against Catalonia in October.

In uttering (1a), its author is doing two things. On the one hand, she is introducing a new event entity in the discourse, that of Red Sox fans cheering the players at a particular time and place. At the same time, she is presenting this event as corresponding to a fact in the world. On the other hand, the author of (1b) is presenting another particular event (the US soccer team playing against the Catalan team in October) and characterizing it only as a mere possibility.

Every time we refer as speakers to a particular situation, we also color it with a specific degree of factuality. We can present it as an unquestionable fact, or express some degree of uncertainty if we are not sure whether the situation holds, or will hold, in the world. This last issue, the factuality status of eventualities mentioned in discourse, is the focus of the present work.

2.2 Defining event factuality

For the current discussion, we define event factuality as the level of information expressing the commitment of relevant sources towards the factual nature of events mentioned in discourse.[2] Events are couched in terms of a veridicality axis that ranges from truly factual to counterfactual, passing through a spectrum of degrees of certainty. In some contexts, the factual status of events is presented with absolute certainty. Depending on the polarity, events are then depicted as either *facts* (2) or *counterfacts* (3).

(2) Five U.N. inspection teams visited a total of nine other sites.
(3) The size of the contingent was not disclosed.

In other contexts, events are qualified with different shades of uncertainty. Combining that with polarity, events will be presented as *possibly factual* (4) or *possibly counterfactual* (5).

(4) United States may extend its naval quarantine to Jordan's Red Sea port of Aqaba.
(5) They may not have enthused him for their particular brand of political idealism.

Factuality can therefore be characterized by means of a double-axis scale. One is the axis of polarity, which defines a binary distinction: positive versus negative. The other, that of certainty, is expressed as a continuous scale. The resulting system is illustrated in Fig. 1.

2.3 Related notions

As defined, the axis of certainty configuring one of the facets of event factuality is related to epistemic modality (a category standardly accepted in the literature) and, in a more secondarily way, to the system of evidentiality.

[2] In this article, the term *event* will be used in a very broad sense to refer to both processes and states, but also other abstract objects such as propositions, facts, possibilities, etc.

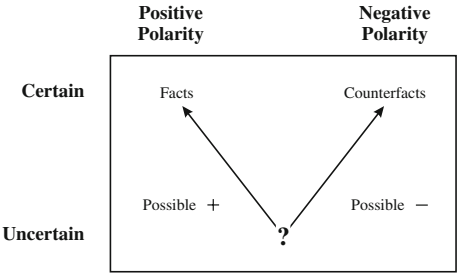

Fig. 1 The double range of factuality

2.3.1 Epistemic modality

The phenomenon of epistemic modality has been studied from both logical and linguistics traditions. Although each discipline adopts a different perspective, in both cases the notion is clearly related to the concept of event factuality.

Work within the logical tradition has distinguished among different kinds of modality: *alethic*, which has to do with necessary and contingent truth of propositions, *deontic* (sometimes also named *root modality*), concerned with obligations and permissions), and *volitional, or bouletic, modality*, which deals with speaker's desires. However, modality in natural language is basically epistemic in nature (Lyons 1977; Kiefer 1987), the type that concerns us here.

Within linguistics, authors from different traditions converge in analyzing modality as a subjective component of discourse, a view that is adopted in the present analysis. Lyons (1977), for instance, presents epistemic modality in terms of a speaker's commitment; Chafe (1986) defines it as the degree of reliability of a situation being a fact as assessed by authors; according to Palmer (1986), it expresses the degree of commitment of the author towards the certainty of an uttered proposition; and Kiefer (1987) claims that modal expressions in natural language generally express the speaker's attitude toward a situation.[3]

2.3.2 Evidentiality

We acquire knowledge about situations in the world through very varied means, from the most physical, direct way, to purely speculative processes. Then, when talking, we can refer to the way we learned about those situations in order to strengthen or qualify our assertions about them. Some languages grammaticalize this into the system known as evidentiality. Evidentiality is thus concerned with the origin of information or, in other words, the way in which information is acquired (Van Valin and LaPolla 1997).

Event factuality is related to evidentiality in that different types of evidence seem to have an effect on the way the factuality of an event is ultimately evaluated. For example,

[3] This is distinct from most of the work within truth-conditional semantics, which conceives of modality as independent from the speaker's perspective (e.g., Kratzer 1991).

the use of the perception predicate *see* in (6a), in contrast to the inference predicate *deduce* in (6b), expresses that the source assessing the factual status of the underlined events in the first example has a stronger evidence for judging them as facts in the world than the source assessing the factuality of the underlined event in the second example.[4]

(6) a. He **saw** the gunman standing and firing, with a straight arm toward the counter.
b. Amiroutzes, the Emperor's Chancellor, **deduced** that the prince Uzum Hasan wished to rear a child of the blood to lead armies one day against Trebizond.

Recent research in this area agrees that English has no overt grammatical system of evidentiality and expresses evidential-related knowledge by means of various strategies, e.g., adverbials like *reportedly* or constructions involving complementation (de Haan 2000; Mushin 2001; Aikhenvald 2004). A common feature of these strategies in English is that they tend to incorporate an epistemic extension as well. For example, we understand events qualified by a predication of inference (*I deduce*) as less certain than those qualified by a predication of direct perception (*I saw*).

2.4 Linguistic means of expressing factuality

Event factuality is conveyed by means of explicit polarity and modality markers, which in natural languages tend to participate in well-defined grammatical systems (e.g., affixes, clitics, or verbal auxiliaries). As it happens, however, there are many other mechanisms conveying distinctions at this level, including lexical items, morphological elements, syntactic constructions, and discourse relations between clauses or sentences.[5]

The present section examines in some detail the linguistic means used to express event factuality in English, which is the only language currently represented in FactBank. However, the information presented here is easily applied to other languages, such as the Romance and Germanic language families.

2.4.1 Polarity particles

Polarity particles express the positive or negative factuality of events mentioned in text. These include elements of a varied nature: adverbs (*not*), quantifiers (*no, none*), pronouns (*nobody*), etc. Unless they are combined with other elements, such as modality operators, polarity particles leave no room for degrees of uncertainty and only switch the original polarity of its context.

Polarity particles can be introduced at different structural levels: at the clausal level of the event, i.e., immediately scoping over the event-referring expression

[4] Here and throughout the rest of the article, events in the examples will be identified by marking only their verb, noun, or adjective head, together with polarity particles and auxiliaries when deemed necessary. This follows the convention assumed in TimeML, the specification language used to represent event and temporal information in the corpus presented here (Pustejovsky et al. 2006).

[5] Some authors use the term *hedging* to refer to markers of modality expressing the degree of commitment of the source towards the certainty of a proposition. See, e.g., Clemen (1997).

(7a); at the subclausal level, affecting one of the arguments of the event (7b–c); or at the lexical level, e.g., by means of affixes such as *un-* in *unable* (Huddleston 1984; Quirk et al. 1985). In the examples below, the negative particles are underlined whereas the negated events are in bold face.

(7) a. She didn't **follow** the rules.
b. Neither proposal **was** satisfactory.
c. The two teenagers **went** nowhere.

It should be pointed out that the presence of a negative polarity marker does not always result in the negation of the event it scopes over (cf. Polanyi and Zaenen 2005). For example, in (8) event *e* is evaluated as positive because the polarity marker scoping over it, *not*, is in turn affected by the negative polarity conveyed by *denied*.

(8) The prosecutor **denied** that Spelke did **not** open$_e$ the door.

2.4.2 *Modality particles*

Modality particles contribute different degrees of certainty to a given event. In English, the main sources of modality information belong to different parts of speech, namely, verbal auxiliaries, adverbials, and adjectives. Each of these categories displays an equivalent gradation of modality, in the sense that each element in one category has a semantic correspondent to the elements in the other (Givón 1993). This is illustrated in Table 1 below for particles expressing epistemic modality.

Other types of modality play a role as well in characterizing the factual nature of events in text. For example, events that are modalized as deontically possible (i.e., allowed, permitted) or deontically necessary (i.e., required) are presented as uncertain. Deontic modalities have a future-projecting effect on the modalized event. The fact that an event is allowed or required does not necessarily imply that it will ever take place in the world.

2.4.3 *Event-selecting predicates (ESPs)*

In many cases, the factuality of events is conveyed by what we refer to as event-selecting predicates (ESPs). These are predicates (either verbal, nominal, or adjectival) that select for an argument denoting an event of some sort. Syntactically, they subcategorize for a *that-*, gerundive, or infinitival clause, but also for NPs

Table 1 Equivalent gradation of modality among different categories

	Modal auxiliaries	Modal adverbs	Modal adjectives
Possible	could, may	perhaps, maybe	it is possible that…
Probable	should, will	probably	it is likely/probable that…
Certain	must, have to	necessarily, certainly	it is certain/impossible that…

headed by event-denoting nouns. Examples of ESPs are verbs like *claim, suggest, promise, offer, avoid, try, delay, think*, nouns like *promise, hope, love, request*, and adjectives such as *ready, eager, able*. The ESPs in (9) are in bold face, while their embedded events are underlined.

(9) a. The Human Rights Committee **regretted** that discrimination against women persisted in practice.
b. Uri Lubrani **suggested** Israel was **willing** to withdraw from southern Lebanon.
c. Kidnappers **kept** their **promise** to kill a store owner they took hostage.

ESPs are of interest here because they project factuality information on the event denoted by its argument through syntactic means. In other words, they qualify the degree of factuality of that event, which then can be marked as:

– **Factual**. The event is presented as corresponding to a fact in the world. This is the case of complements of: certain implicative predicates (10b) (Karttunen 1970); factive predicates (10b) (Kiparsky and Kiparsky 1970); perception predicates (10c); aspectual predicates (10d), change-of-state predicates (10e), etc.

(10) a. Russian spies **managed** [to intercept fax communications of top U.S. defense contractors].
b. Furrow's neighbors **knew** that [he was a neo-Nazi].
c. Austin **saw** a man [get shaved in Oxford].
d. Culturebox just **finished** [reading one of the pieces of evidence for the case].
e. Japan saved itself by dramatically **increasing** [its exports].

– **Counterfactual**. The event does not correlate to any situation in the world. For example, arguments of certain implicative predicates, like *avoid* and *prevent* (Karttunen 1970; Condoravdi et al. 2001; Koenig and Davis 2001).
– **Not totally certain** (i.e., possible or probable). It is not certain whether the event took or will take place in the world. This is the case of the complements to the so-called weak assertive predicates, such as *think* and *believe*, mainly when used with a first person subject and in present tense (11a) (Hooper 1975). Other predicates (e.g., *speculate, suspect*) bear this nuance of uncertainty as well, regardless of tense and grammatical person in the subject (11b). They have been acknowledged by many authors, e.g., Bach and Harnish (1979), Wierzbicka (1987), and Dor (1995).

(11) a. I **think** [he wants to hire a woman].
b. The WSJ editorial page **speculated** that [the president suffers from an anti-social personality disorder].

– **Underspecified**. The event is mentioned in discourse, but there is no information provided about its factual status. There are several predicate classes which create such an effect. For example, volition (e.g., *want, wish, hope*), commitment (*commit, offer, propose*), and inclination predicates (*willing, ready, eager, reluctant*), among others. All of them select for an argument of what Asher (1993) calls projective proposition type.

In addition to the text author, some ESPs present one or more sources that are also committing to the factuality of their embedded event. This is observed with reporting predicates (e.g., Bergler 1992; Dor 1995; Waugh 1995). The event argument of verbs like *affirm* or *say*, for example, is presented as true to the subject of the verb but not according to the text author, who remains uncommitted. Other classes of verbs introducing an additional source are knowledge and belief predicates. Because of the relevance of sources in the attribution of factuality values to events, we distinguish here between two kinds of ESPs: those that introduce an additional source as a factuality evaluator in the discourse, and those that do not. They are distinguished below.

Source introducing predicates (SIPs). These are ESPs contributing an additional source relative to which the factuality of the embedded event is assessed. Consider the sentences in (12). They each present an SIP, *suspects* and *knows* (in bold face), which introduces the participant *Berven* as a relevant source for computing the factuality of event *e* (Freidin leaving the country in June). The use of *suspects* in (12a) presents Berven as considering *e* to be just a possibility, whereas in (12b) she is depicted as considering it a fact. Moreover, the text author adopts a particular view as well. He is uncommitted in the first sentence but agrees with Berven's judgement, in the second.

(12) a. Berven **suspects** that Freidin $\underline{\text{left}}_e$ the country in June.
b. Berven **knows** that Freidin $\underline{\text{left}}_e$ the country in June.

SIPs can be classified as belonging to one of the following classes:

- Predicates of report: for example, *say, tell, add; claim, argue*—even if they express report by means other than oral; e.g., *write, publish, post*.
- Predicates of knowledge: they can express the state of having knowledge (*know, understand, remember*), acquiring knowledge (*learn, find out, discover*), losing knowledge (*forget*), admitting knowledge (*acknowledge, accept, admit*).
- Predicates of belief and opinion: such as *think, consider, guess, predict, suggest*.
- Predicates of doubt: such as *doubt, wonder, ask*. They generally subcategorize for a *wh*-clause complement.
- Predicates of perception: e.g., *see, hear, feel*.
- Predicates expressing proof: e.g., *prove, show, support, explain*.
- Predicates expressing some kind of inferencing process: e.g., *infer, deduce; conclude, decide (that); mean, seem, appear* (as in: *it means/appears/seems that*).
- Predicates expressing some psychological reaction as a result of an event or situation taking place: *regret, be glad/pleased (that), like (that), love (that)*.

The new sources correspond therefore to agents of speech acts, holders of opinions, experiencers of psychological reactions, etc.

Non-source introducing predicates (NSIPs). These are ESPs not contributing any additional source. The only source at play here is the text author—unless additional sources have been previously introduced. NSIPs include a varied set of classes, among which are the following:

- Implicative and semi-implicative predicates: such as *fail, manage*, or *allow*.
- Predicates introducing a future event as their complement: this type includes many different semantic classes, such as volition (*want*), commissive (*offer*), and command (*require*) predicates. The complement event is always of an intensional nature. Hence, its factuality value is uncertain.
- Change of state predicates: for example, *increase*, *change*, or *improve*.
- Aspectual predicates: such as *begin*, *continue*, and *terminate*.

It has been argued that some of the predicates listed above (mainly, those in the second bullet (e.g., *want*, *offer*, or *require*) also introduce a new source in discourse. They are in fact treated in this way within the MPQA corpus. Here, however, we consider them as NSIPs. Even though they have an epistemic component in the sense that they characterize the factuality of the their embedded event as uncertain, there is a difference in the core meaning of SIPs vs. this second type of predicates. SIPs depict the attitude of their subject as being epistemic in nature; that is, as concerning the degree of certainty that an event has taken (or will take) place in the world. On the other hand, predicates like *want* or *offer* focus on the role of their subjects as either having some degree of responsibility on the potential event (e.g., *agree/promise/offer to go; force somebody to go*), or being in a greater or lesser favorable state towards the accomplishment of the embedded event (e.g., *need/want to go*). In other words, SIPs are predicates concerning epistemic modality, whereas the NSIPs discussed here express distinctions within the space of deontic modality.

2.4.4 Syntactic constructions

Some syntactic constructions involving subordination introduce factuality information of some sort. In some cases, the embedded event is presupposed as holding as fact; e.g., non-restrictive relative clauses (13a) and cleft sentences (13b).

(13) a. Rice, who became secretary of state two months ago today, took stock of a period of tumultuous change.
b. It was Mr. Bryant who, on July 19, 2001, asked Rep. Bartlett to pen a letter to him.

In others, the event denoted by the embedded clause is intensional in nature, therefore presented as underspecified with respect to its factuality status; e.g., rationale clauses (14a) and conditional constructions (14b).

(14) a. The environmental commission has adopted regulations to ensure that people are not exposed to radioactive waste.
b. EZLN will return to the negotiating table if the conflict zone is **demilitarized**.

2.4.5 Discourse structure

An additional means for conveying factuality information is available at the discourse level. Some events may first have their factual status characterized in one

way, but then be presented differently in a subsequent sentence. The most common mechanisms responsible for this type of overwriting are relations of opposition between clauses or sentences, expressed by means of discourse connectors. Consider the following example, concerning the event of drug dealers being tipped off (underlined):

(15) 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.

Strictly speaking, discourse (or rhetorical) relations are not factuality markers per se, but contribute to the final assessment of the factuality of an event according to a given source. Despite their pervasiveness, the present work will ignore factuality assessments expressed (or overwritten) at an inter-sentential discourse (or even cross-document) level.

3 Challenges in identifying event factuality

Identifying event factuality in text presents challenges at several levels of analysis. First, if the set of factuality scales is not based on solid criteria, we may end up with inconsistently annotated data, which will be of little use for informing automatic identification tasks. Second, factuality is in many cases the result of the interaction of several different factuality markers. They may all be within the local context of the event, but it is also common for them to be at different levels. And third, the factuality of an event is always relative to at least one source, but often times more than one as well, all of which must be included as part of the annotation. The following subsections elaborate on these issues.[6]

3.1 Distinguishing among factuality degrees

Given the number and diversity of factuality markers available in natural languages, speakers can choose to be very precise in expressing different shades of factuality. For example, assuming only a limited number of words in English, one can create the following distinctions: *improbable, slightly possible, possible, fairly possible, probable, very probable, most probably, most certainly, certainly*. As agreed by many linguists, modality in natural language is a continuous category, but speakers are able to map areas of this axis into discrete values (Lyons 1977; Horn 1989; de Haan 1997). The issue is therefore identifying the factuality distinctions we tend to use in everyday language.

Our goal is to obtain a consistent annotation while achieving a satisfactory description of the phenomenon. Thus, we need to ensure (a) that these distinctions reflect our linguistic intuitions as speakers, and (b) that it is possible to define a set

[6] See Saurí (2008) for a more comprehensive view on the factuality of events and its identification.

of sound and stable criteria for differentiating among them. The factual value of markers such as *possibly* and *probably* is fairly transparent. What, however, is the contribution of elements like *think*, *predict*, *suggest* or *seem*?

3.2 Interaction between factuality markers

The set of criteria mentioned above will, in addition, help discern the resulting factuality value in cases where two or more factuality markers interact among them. In (16), for example, event *e* (underlined) is under the scope of markers *suggested* and *must*. The latter seems to convey a stronger degree than the former. What, then, is the resulting degree of factuality characterizing event *e*?

(16) But at the time, characteristically, it was **suggested** that the book **must** have been written$_e$ by Jean-Paul Sartre.

The following examples illustrate the issue of interacting factuality markers in more detail.

(17) a. The Royal Family will **continue** to **allow** detailed fire brigade **inspections**$_e$ of their private quarters.
b. The Royal Family will **continue** to **refuse** to **allow** detailed fire brigade **inspections**$_e$ of their private quarters.
c. The Royal Family **may refuse** to **allow** detailed fire brigade **inspections**$_e$ of their private quarters.[7]

In all three examples, the event *inspections* is directly embedded under the verb *allow*, which is generally used as a two-way implicative predicate, that is, as a predicate that holds a direct correlation between its truth (or falsity) and that of its embedded event (Karttunen 1970). Note that in each of the sentences, the event *inspections* receive a different interpretation depending on the elements scoping over *allow*. For instance, in (17a), where *allow* is embedded under the factive predicate *continue*, *inspections* is characterized as a fact in the world; i.e., there have been such inspections. Example (17b), on the other hand, depicts *inspections* as a counterfact because of the effect of the predicate *refuse* scoping over *allow*. Now contrast the two previous sentences with that in (17c), where the factual status of the event *inspections* is uncertain due to the modal auxiliary *may* scoping over *refuse*.

Hence, the factuality status of a given event cannot be determined from the strictly local modality and polarity operators scoping over that event alone; rather, if present, other non-local markers must be considered as well to arrive at an interpretation.[8] Annotating factuality with a shallow, feature-based approach, without considering the interaction and scope of the various markers would therefore miss an important piece of information.

[7] The original sentence in this set is (17b), from the British National Corpus.

[8] Furthermore, Nairn et al. (2006), Saurí and Pustejovsky (2007), and Saurí (2008) show that the interaction among all these elements can be modeled in a predictable way.

3.3 Relevant sources

The third challenge to encoding event factuality involves the notion of perspective. Different discourse participants may present divergent views about the factuality nature of the very same event. Recognizing these sources is crucial for any task involving text entailment, such as QA or narrative understanding. For example, event *e* in (18) (i.e., Slobodan Milosevic having been murdered in The Hague) will be inferred as a fact in the world if it cannot be qualified as having been asserted by a specific source, namely, Milosevic's son.

(18) Slobodan Milosevic's son said Tuesday that the former Yugoslav president had been **murdered**$_e$ at the detention center of the UN war crimes tribunal in The Hague.

By default, events mentioned in discourse always have an implicit source, viz., the author of the text. Additional sources are introduced in discourse by means of Source-Introducing Predicates (SIPs)—*cf.* Sect. 2.4.3.

The status of the additional sources is, however, different from that of the author of the text. For instance, in (3) the reader learns Izvestiya's position only according to what the author asserts. In other words, the reader does not have direct access to the factual assessment of Izvestiya about event e_2—or, equivalently, to the assessment of G-7 leaders about e_3.

(19) Izvestiya **said**$_{e1}$ that the G-7 leaders **pretended**$_{e2}$ everything **was OK**$_{e3}$ in Russia's economy.

Thus, we need to appeal to the notion of *nested source* as presented in Wiebe et al. (2005). That is, *Izvestiya* is not a legal or licensed source of the factuality of event e_2, but *Izvestiya* according to the author, represented here as *izvestiya_author*.[9] Similarly, the source referred to by the G-7 leaders corresponds to the chain: *g7leaders_izvestiya_author*.

In some cases, the different sources that are relevant for a given event may coincide with respect to its factual status, but in others, they may be in disagreement. In (19), for example, event e_3 is assessed as being a fact according to the G-7 leaders (corresponding to source *g7leaders_izvestiya_author*), but as being a counterfact according to Izvestiya (i.e., *izvestiya_author*). The text author, on the other hand, remains uncommitted.

To sum up, the factuality value assigned to events in text must be relative to the relevant sources at play in the discourse, of which there may be many. Only under this assumption is it possible to account for the potential divergence of opinions about the factuality status of events (e.g., Milosevic having been murdered), as is common in news reports.

[9] This is equivalent to the notation < *author*,*izvestiya* > in Wiebe's work. Here, we adopt a reversed representation of the nesting (i.e., the non-embedded source last) because it positions the most direct source of the event at the outmost layer, thus facilitating its reading.

4 Factuality information in existing corpora

The various information structures relating to factuality (e.g., modality, evidentiality) have been widely discussed in the fields of linguistics and philosophy but have received less attention in NLP. Only recently there has been some significant work in the field towards representational frameworks capable of accounting for these notions.

The work closest to ours is the research carried out by Rubin (2006, 2007), which is concerned with the notion of *certainty*, and explores it by means of an annotation experiment on written news discourse. *Certainty* there is understood as "a type of subjective information available in texts and a form of epistemic modality expressed through explicitly-coded linguistic means" (Rubin et al. 2005, p. 5). Certainty is therefore conceived along the same terms as the axis of modality that configures the system of event factuality, presented in Sect. 2.2.

Rubin's approach is interesting in that it is both data-driven and framed within an NLP perspective. Furthermore, it resulted in the creation of a corpus of 80 documents (news reports and editorials) with explicit certainty markers manually annotated. Nevertheless, it suffers from a lack of precision in characterizing the phenomenon. First, "certainty" is defined as applying to statements but, as it happens, one statement my express more than one event (or proposition), each of them qualified with a different degree of certainty. Consider example (20), where the event denoted by the relative clause (underlined) is presented with total certainty, whereas that expressed by the main predication (in italics) is depicted as a possibility in the real world.[10]

(20) In future primaries, where crossover voting is barred, *Bush may well have an easier time*.

Second, the distinctions among different degrees of certainty is not based on a set of stable semantic criteria, which in our opinion partly explains the low interannotation agreement scores obtained in two experiments on assigning certainty degrees to statements ($\kappa_{cohen} = 0.15$ and $\kappa_{cohen} = 0.41$). And finally, it does not deal with the possible interaction between markers, nor with the potential presence of multiple sources.

In addition to this work, factuality-related information is annotated in at least five corpora, although it must be pointed out that factuality is not the main component targeted by any of them. These corpora are: the MPQA Opinion Corpus (Wiebe et al. 2005), the Penn Discourse TreeBank (Miltsakaki et al. 2004), TimeBank (Pustejovsky et al. 2006), different versions of the ACE corpus for the Event and Relation recognition tasks (see, e.g., ACE 2008), and finally the corpus created by Bethard et al. (2004) in order to explore the automatic extraction of opinion propositions and their sources.

The factuality-relevant expressions annotated in the MPQA Opinion Corpus are private states (opinions, beliefs, thoughts) and speech events. They both convey the stance of a source with regard to what is believed or said. Nevertheless, event

[10] From Rubin (2006, p. 59).

factuality is not the focus of the annotation, and hence these events and states are not characterized in terms of the factual degree they convey but in terms of attitude; i.e., objective (non-affectual state) vs. subjective (positive or negative).

One feature that both the MPQA Opinion Corpus schema and our model of event factuality share is the encoding of sources. Both approaches structure them as chains of nested sources. For our task, however, the MPQA Opinion Corpus is limited in that in most cases, it only acknowledges one relevant source for each event. That is, once a new source is introduced in discourse (e.g., source: *milosevic'sSon_author*, in example (18) above), the previous sources (here, only one: *author*) will not be considered as relevant for evaluating the embedded events. There is only a small number of cases in which previous sources are kept as relevant for subsequent events, namely, those in which the new source is introduced by a predicate like *pretend* or *slander*, which explicitely express a disagreement between two sources.

The corpus created by Bethard et al. (2004) is similar to the previous one in that it aims at encoding opinions and their holders. However, it is even simpler in that it does not appeal to the concept of nested source, nor does it classify opinions along either a factuality or attitude-based axis.

The Penn Discourse TreeBank (PDTB) seems closer to our perspective in that it contemplates the attribution of abstract objects (corresponding here to what we refer to as events), and encodes both their sources and the degree of factuality associated with them (Prasad et al. 2007). In spite of these similarities, there are two significant differences. With regard to sources, PDTB neither encodes the nesting relation that can hold among them, nor accounts for the possibility of several sources for a given abstract object (or event). The second difference concerns the factuality degree associated with the attributed event, which is assigned based on the type of predicate embedding it. In particular, events embedded under communication predicates (e.g., *say*) are characterized as asserted; events embedded by propositional attitude predicates (e.g., *think*), as beliefs; and events embedded under factive predicates (e.g., *know*), as facts.

As it happens, however, each of these types of predicates is not uniform in terms of the factuality they project to the embedded event. *Suggest*, for instance, is a communication verb which nevertheless conveys nuances of belief or uncertainty (e.g., *It has been suggested, for example, that mammals took over from the dinosaurs about 65 million years ago*.). Similarly, *forget* is a factive predicate which, contrary to others in its class, expresses an uncommitted (or ignorant) stance of the cognizer (i.e., the participant expressed by its subject) with regards to the factual status of its embedded complement. The classification misses therefore important factuality distinctions. The PDTB annotation, in addition, has the feature Determinacy, with the binary distinction *Null* and *Indet*, expressing the factual and non-factual status of events, respectively. This is, however, a very basic distinction which lacks the different degrees of possibility used by speakers to characterize events.

Factuality information in ACE is presented in terms of modality. It distinguishes between *asserted* (for situations which can be interpreted as pertaining to "the real world") and *other* (for situations holding in "a particular counterfactual world"). Like in the case of PDTB, this is a very simple classification which seems to miss basic distinctions. Furthermore, assessing the modality of events and relations is performed independent of the relevant sources in each case.

The last corpus to evaluate is TimeBank, a corpus annotated with TimeML (Pustejovsky et al. 2005), a specification language representing temporal and event information in text. The factuality-relevant information encoded in TimeBank is mainly lexical: grammatical particles expressing event modality and polarity, as well as event-selecting predicates (i.e., ESPs—*cf.* Sect. 2.4.3), which project a factual value to their embedded event by means of subordination links (or SLINKs). Thus, TimeBank provides us with some of the basic components expressing factuality information in text—which is a consequence of the explicit surface-based approach of TimeML. But whereas there is some characterization of event factuality (through SLINKs), it does not deal with the interaction among the different markers scoping over the same event.

5 Towards a corpus of event factuality

Building a corpus annotated with event information and their factuality values serves two purposes. On the one hand, it tests the designed model of event factuality. On the other hand, it provides the community with a data resource that can be of great help for further analyzing the phenomenon, as well as developing, training, and testing tools for its automatic identification.

In this section, we present FactBank, a corpus designed to encode event factuality in natural language text. We first focus on the two main decisions concerning its creation, namely, setting the factuality values used for its annotation, and selecting the documents. Because FactBank consists of the same set of documents as TimeBank, it can be seen as adding an additional layer of semantic information to that corpus. The final two parts of the current section will address the mapping between these two corpora.

5.1 Defining the descriptive framework

We argued above that factuality is a kind of information that is difficult to annotate, mainly because it expresses distinctions along a non-discrete system; i.e., that of epistemic modality. According to many linguists, however, speakers tend to map areas of that system into discrete values (Lyons 1977; Horn 1989; de Haan 1997). In order to obtain consistent annotation that can be used for informing systems of automatic extraction, it is therefore fundamental that we establish (a) a discrete set of factuality values which effectively reflect the main distinctions applied in natural languages; and (b) a battery of sound criteria that allow annotators to differentiate among these values. Only in this way is it possible to obtain an acceptable degree of interannotation agreement—and thus, of annotation consistency. The next subsections focus on each of these two targets.

5.1.1 Set of factuality values

The set of values for characterizing event factuality must account for distinctions along both axes of polarity and modality. While polarity is a binary system with the

values *positive* and *negative*, epistemic modality constitutes a continuum ranging from *uncertain* (or *possible*) to absolutely *certain* (or *necessary*). We therefore need a discrete categorization of that second system.

Within modal logic two operators are typically used to express a modal context: necessity (□) and possibility (◇). On the other hand, most of the work in linguistics points towards a three-fold distinction: *certain*, *probable*, and *possible* (e.g., Lyons 1977; Halliday and Matthiessen 2004). Interestingly, Horn (1989) analyzes modality and its interaction with polarity based on both linguistic tests and logical relations at the basis of the Aristotelian Square of Opposition. Although he agrees that modality is a continuous category, this view provides a good grounding for differentiating the three major modality degrees just mentioned.

In Horn's work, the system of epistemic modality is analyzed as a particular instantiation of scalar predication.[11] The relations holding among predicates of the same scalar predication are manifested in syntactic contexts like the following (Horn 1972):

– Contexts in which the speaker is explicitly leaving the possibility open that a higher value on the relevant scale obtains:

  1. (at least) P_{n-1}, if not (downright) P_n.
  2. P_{n-1}, {or/ and possibly} even P_n.
  3. not even P_{n-1}, {let alone/ much less} P_n.

– Contexts in which the speaker asserts that a higher value in the scale is known to obtain:

  1. P_{n-1}, {indeed/ in fact/ and what is more} P_n.
  2. not only P_{n-1} but P_n

In particular, he proposes the epistemic modal scale of: <*certain*, {*probable/ likely*}, *possible*>. The appropriateness of this scale can be checked based on the tests above. The symbol # is used to express semantic anomaly.

(21) a. possible, if not likely — #likely, if not possible
b. likely, or even certain — #certain, or even likely
c. possible, and in fact likely — #likely, and in fact possible

The same tests allow Horn to conclude that the elements in the negative counterpart are ranked as <*impossible*, *unlikely*/*improbable*, *uncertain*> (22), and must constitute an independent scale by themselves since they cannot be copredicated with elements in the positive scale (23).

(22) a. possibly not, if not certainly not — #certainly not, if not possibly not
b. possibly not, or even certainly not — #certainly not, or even possibly not
c. possibly not, and in fact certainly not — #certainly not, and in fact possibly not

[11] Scalar predications are conceived as collections of predicates P_n such as $<P_j, P_{j-1}, \ldots, P_2, P_1>$, where P_n outranks (i.e., is stronger than) P_{n-1} on the relevant scale.

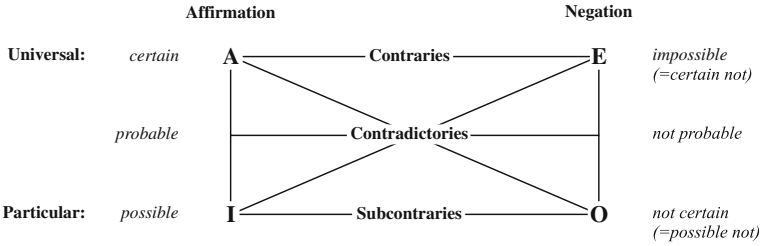

Fig. 2 SO for epistemic modals, adapted from Horn (1989, p. 325)

(23) a. #possibly not, if not likely — #likely if not possibly not
b. #likely, or even certainly not — #certainly not, or even likely
c. #possibly not, and in fact likely — #likely, and in fact possibly not

In sum, there are two epistemic scales which differ in quality (positive vs. negative polarity):

(24) a. ⟨*certain*, *likely* (*probable*), *possible*⟩
b. ⟨*impossible*, *unlikely* (*improbable*), *uncertain*⟩

The beauty of the system can be appreciated when mapped to the traditional Square of Opposition (SO), employed to account for the interaction between negation and quantifiers or modal operators (Horn 1989, following Aristotle). Its basic structure (applied to epistemic modals) is shown in Fig. 2.

On the horizontal axis, we have a distinction in quality: positive versus negative polarity.[12] On the other hand, the vertical axis represents a difference in quantity: universal versus particular. The epistemic modal operators are displayed in italics. This version of the SO also adds the intermediate values of the epistemic scale (*probable* and *not probable*), as proposed by Horn.

The Square of Opposition allows us to illustrate the logical relations holding between two operators paired at the horizontal axis. Pairs A/O, I/E, and the one with the two mid values are *contradictories*. Contradictories satisfy both the Law of Contradiction (LC), which states that a statement cannot be true and false at the same time, and the Law of Excluded Middle (LEM), which states that a statement must be either true or false. On the other hand, A/E are *contraries*: they satisfy the LC but not the LEM, since both can be false at the same time. Finally, I/O are *subcontraries*: both values can hold at the same time. The following examples illustrate it with the pairs at the low, mid, and high level:

(25) SUBCONTRARY: *possible, possible not*
a. Not satisfying LC: *It is possible that P and it is possible that not P.*
b. Satisfying LEM: *#It is neither possible that P nor possible that not P.*

[12] The vowels naming the vertices, which are derived from Latin verbs **a***ffi***r***mo* 'I affirm', and *n***e***g***o** 'I deny', reflect this distinction.

Table 2 Factuality values

	Positive (+)	Negative (−)	Underspecified (u)
Certain (CT)	Fact: <CT,+>	Counterfact: <CT,−>	Certain but unknown output: <CT,u>
Probable (PR)	Probable: <PR,+>	Not probable: <PR,−>	(NA)
Possible (PS)	Possible: <PS,+>	Not certain: <PS,−>	(NA)
Underspecified (U)	(NA)	(NA)	Unknown or uncommitted: <U,u>

(26) CONTRADICTORY: *likely, likely not*
  a. Satisfying LC: *#It is likely that P and it is likely that not P.*
  b. Satisfying LEM: *#It is neither likely that P nor likely that not P.*

(27) CONTRARY: *certain, certain not*
  a. Satisfying LC: *#It is certain that P and it is certain that not P.*
  b. Not satisfying LEM: *It is neither certain that P nor certain that not P.*

Based on Horn's distinctions, we can represent degrees of factuality by means of the features in Table 2, where the factuality value of events is characterized as the pair <*mod*, *pol*>, containing a modality and a polarity value.[13]

The polarity axis divides into *positive* (+), *negative* (−), and *underspecified* (u), while the modality axis distinguishes among *certain* (CT), *probable* (PR), *possible* (PS), and *underspecified* (U). The *underspecified* values on both axes are added to account for cases of non-commitment of the source or in which the value is not known.

The table includes six fully committed (or specified) values (<CT,+>, <CT,−> , <PR,+>, <PR,−>, <PS,+>, <PS,−>), and two underspecified ones: the partially underspecified <CT,u>, and the fully underspecified <U, u>. The use of each of them is summarized in what follows. From here onwards, they will be represented in the abbreviated form of CT+, PR−, Uu, etc.

(28) **Committed Values:**

CT+ According to the source, it is **certainly** the case that X.
PR+ According to the source, it is **probably** the case that X.
PS+ According to the source, it is **possibly** the case that X.
CT− According to the source, it is **certainly not** the case that X.
PR− According to the source it is **probably not** the case that X.
PS− According to the source it is **possibly not** the case that X.

(Partially) Uncommitted Values:

CTu The source knows whether it is the case that X or that not X.
Uu The source does not know what is the factual status of the event, or does not commit to it.

The use of the fully committed values should be clear from the paraphrases above, but uncommitted values deserve further explanation. The partially underspecified value CTu is for cases where there is total certainty about the factual nature

[13] Semantically, this can be interpreted as: *Val*(*mod*,*Val*(*pol*,*e*))—i.e., the modal value scopes over the polarity value.

of the event but it is not clear, however, what the output is—e.g., (29). The fully underspecified value Uu, on the other hand, is used when any of the following situations applies: (i) The source does not know what the factual status of the event is, as in (30a); (ii) the source is not aware of the possibility of the event—e.g., (30b); or (iii) the source does not overtly commit to it—e.g., (30c). The following examples illustrate each of these situations for the underlined event when evaluated by source *John*:

(29) **John** knows whether Mary came.
(30) a. **John** does not know whether Mary came.
b. **John** does not know that Mary came.
c. **John** knows that Paul said that Mary came.

5.1.2 Discriminatory tests

In determining the factuality of events, the polarity parameter offers no problem, but distinguishing among modality values (e.g., between *possible* and *probable*) is not always obvious. For that reason, we designed a battery of discriminatory tests of copredication to be used by annotators; that is, the original sentence is conjoined with a second sentence (or clause) where the event in question appears qualified with a different polarity degree and possibly, also, modality. These tests are based on the logical relations used in Horn (1989) to identify the basic degrees of epistemic modality (i.e., Law of Contradiction and Law of Excluded Middle). Note that they are also conceived as relations that conjoin predications of different modality and polarity values. They are the following:

Underspecification (U) versus different degrees of certainty (CT, PR, PS): Events with an underspecified value can be copredicated with both: a context in which they are characterized as certainly happening (CT+), and a context in which they are presented as certainly not happening (CT−). For example, sentence (31) can be continued by either fragment in (35), the first of which maintains the original underlined event as certainly happening (CT+), and the second as certainly not happening (CT−). This is not the case, however, for sentences in (32–34), where the underlined event is explicitly characterized as certain, probable and possible, respectively.

(31) Iraq has agreed to allow Soviets in Kuwait to leave. (Uu)
(32) Soviets in Kuwait will finally leave. (CT+)
(33) Soviets in Kuwait will most probably leave. (PR+)
(34) It is possible that soviets in Kuwait will leave. (PS+)
(35) a. … They will take the plane tomorrow early in the morning. (CT+)
b. … However, most of them decided to remain there. (CT−)

In general, committed degrees of modality (CT, PR, PS) can be copredicated with a context of certainty (CT) as long as they hold the same polarity as the original. We refer to this copredicative context as *context*:$CT_{=}$, where the subindex = indicates that the same polarity as in the original is kept.

On the other hand, the uncommitted value (U) can be copredicated with all contexts of certainty, regardless of whether the polarity is the same as or different from the polarity in the original context.

Subsequently, the test for distinguishing between underspecified (U) versus specified (CT, PR, PS) modality values will be referred to as *test*:$\mathrm{CT}_{\neq}$. Only the underspecified value satisfies it, that is, it can be copredicated with contexts of certainty presenting a different polarity than the original (*context*:$\mathrm{CT}_{\neq}$). The other values (CT, PR, PS) will fail it.

Absolute certainty (CT) versus degrees of uncertainty (PR, PS): Eventualities presented as certain (CT) cannot at the same time be assessed as *possible* (PS) in a context of *opposite polarity* (*context*:$\mathrm{PS}_{\neq}$).

(36) a. Hotels are only thirty (CT+) percent full.
b. #… but it is possible that they aren't (PS−).
(37) a. Nobody believes (CT−) this anymore.
b. #… but it is possible that somebody does (PS+).

On the other hand, eventualities characterized with some degree of uncertainty (PS or PR) allow for it:

(38) a. I *think* it's not going to change (PR−) for a couple of years.
b. … but it *could* happen otherwise. (PS+)
(39) a. It is *possible* that he died (PS+) within weeks or months of his capture.
b. … but it is also possible the kidnappers kept him alive for a while. (PS−)

In (38), the source expressed by the pronoun *I* characterizes the underlined event as PR− by presenting it under the scope of the predicate *think* used in 1st person. The fragment in (38b) can be added without creating any semantic anomaly. A similar situation is presented in (39): the predicate *possible* characterizes the event as PS+, but the additional fragment presents the possibility of things being otherwise. Hence, the test distinguishing between absolute certainty (CT) versus degrees of uncertainty (PR, PS) is *test*:$\mathrm{PS}_{\neq}$. Events presented as certain will fail it, whereas those with some degree of uncertainty will satisfy it.

Probable (PR) versus possible (PS): As seen, both degrees of uncertainty (PR and PS) accept copredication with PS in a context of opposite polarity (*context*:$\mathrm{PS}_{\neq}$). However, only the lowest degree of uncertainty (PS) accepts copredication with PR in a context of opposite polarity (*context*:$\mathrm{PR}_{\neq}$).

(40) a. I *think* it's not going to change (PR−) for a couple of years.
b. #… but it *probably* will. (PR+)
(41) a. It *may* not change (PS−) for a couple of years.
b. … but it most *probably* will. (PR+)

The test distinguishing between these two values is therefore *test*:$\mathrm{PR}_{\neq}$. Value PR fails it but value PS passes it.

Table 3 summarizes the different copredication tests just introduced. The resulting epistemic modality values assigned to events are listed in the rows, while the tests are presented in the columns. We illustrate how these tests are applied in order to identify the factuality value of event *e* (*change*) in the sentence:

Table 3 Tests for discriminating among modality degrees

	test:CT$_{\neq}$	*test*:PR$_{\neq}$	*test*:PS$_{\neq}$
U	ok	ok	ok
PS	#	ok	ok
PR	#	#	ok
CT	#	#	#

(42) I think it's not going to change$_e$.

Due to the negative particle scoping over *e*, we know it is characterized with a negative polarity. The next step is finding its degree of epistemic modality. First, we check whether it is underspecified (U) by applying *test*:CT$_{\neq}$ (43b), which fails. The event has therefore a committed modality degree (CT, PR, or PS). Next, we analyze if the event is characterized as totally certain (CT) by applying *test*:PS$_{\neq}$ (43c).[14] This test is passed, which indicates that the event is qualified with some degree of uncertainty. There are now two candidate values left, PR and PS. *Test*:PR$_{\neq}$ is applied next in order to discriminate between them (43d). Since this test fails, we can conclude that the modality value for event *e* is PR.

(43)
a. **Original:** I *think* it's not going to change$_e$.
b. *test*:**CT**$_{\neq}$: #… but it is certain that it will. [Testing for value U –negative]
c. *test*:**PS**$_{\neq}$: … but it is possible that it will. [Testing for value CT –negative]
d. *test*: **PR**$_{\neq}$: #… but it is probable that it will. [Testing for value PS –negative]

5.2 Corpus selection

FactBank has been annotated with factuality information according to the annotation framework (set of factuality values and discriminatory tests) presented in the previous section. It consists of 208 documents and contains a total of 9,488 manually annotated events. FactBank includes all the documents in TimeBank and a subset of those in the AQUAINT TimeML Corpus (A-TimeML Corpus).[15] The contribution of each of these corpora to FactBank is shown in Table 4.[16]

Both the TimeBank and A-TimeML corpora are annotated with the TimeML language (Pustejovsky et al. 2005), which encodes temporal-related information: identification of time and event expressions, as well as the temporal relations holding between them. TimeML has a highly surface-based approach, and thus it annotates the grammatical items that have a role in the temporal ordering of events, such as tense, aspect, polarity and modality. In addition, TimeML targets lexical and syntactic elements expressing degrees of modality and factivity in text (Saurí et al. 2006a, b). All these components are of relevance for identifying the degree of factuality assigned to events in text.

[14] This step is applied here only for the purpose of illustrating the complete process, although it should be clear just from the meaning of the sentence that the event *change* in the original example is presented with some degree of uncertainty.

[15] http://www.timeml.org/site/timebank/timebank.html.

[16] The figures reported here update those reported in previous work (Saurí 2008; Saurí and Pustejovsky 2008).

Table 4 FactBank sources

	# Documents	# Events
TimeBank	183 (88%)	7935 (83.6%)
A-TimeML Corpus	25 (12%)	1553 (16.4%)
Total	208	9488

5.3 FactBank as a second layer on top of TimeBank

Because the documents constituting FactBank are also annotated with TimeML in other corpora (TimeBank and A-TimeML), they are now annotated with two levels of factuality information. While TimeBank and A-TimeML encode the structural elements expressing factuality in language, FactBank represents the interpretation that results from the interaction among these elements. Thus, FactBank incorporates a second layer of factuality information on top of that contributed by the TimeML annotation. Consider the following sentence:

(44) Newspaper reports have **said** Amir was infatuated with Har-Shefi and **may** have been **trying** to impress her by killing the prime minister.

Event expressions are underlined, whereas factuality markers are presented in bold face. These are: the ESPs *said*, which affects the factuality nature of its embedded events (*infatuated* and *trying*), and *trying*, which qualifies *impress*. Moreover, the modal auxiliary *may* marks the event expressed by *trying* with a degree of possibility. This information is annotated in TimeML as shown:[17]

(45)
```
Newspaper reports have
<EVENT eid="e22" class="REPORTING" tense="PRESENT" aspect="PERFECTIVE">
said </EVENT>
Amir was
<EVENT eid="e23" class="STATE" tense="PAST">
infatuated </EVENT>
with Har-Shefi and may have been
<EVENT eid="e24" class="I_ACTION" modality="may" tense="NONE" aspect="PERF_PROG">
trying </EVENT>
to
<EVENT eid="e25" class="OCCURRENCE" tense="INFINITIVE">
impress </EVENT>
her by
<EVENT eid="e26" class="OCCURRENCE" tense="PRESPART" aspect="NONE">
killing </EVENT> the prime minister.
<SLINK lid="l50" eventId="e22" subordinatedEventId="e23" relType="EVIDENTIAL"/>
<SLINK lid="l51" eventId="e22" subordinatedEventId="e24" relType="EVIDENTIAL"/>
<SLINK lid="l52" eventId="e24" subordinatedEventId="e25" relType="MODAL"/>
```

[17] TimeML has moved towards a stand-off annotation. The example here is embedded for illustration purposes.

That is, the modality incorporated by *may* to *trying* is annotated in the attribute modality of the EVENT tag of the later. On the other hand, the effect of the ESPs on their subordinated events is expressed by means of SLINKs (for subordination links).

Note however that the factuality information provided is minimal. First, the modality value contributed by the modal auxiliary is not interpreted. The modality attribute encodes only the modal particle string. Second, the events *infatuated* and *trying* are presented as reported by means of the two SLINKs (which are of EVIDENTIAL type). However, the annotation does not provide information concerning the reporting source: who is it, and what is its stance with regards to the factuality status of the embedded event; e.g., events reported as *confirmed* vs. events reported as *suggested* are not characterized with different factuality degrees. Finally, note that neither *said* nor *killing* are annotated with any kind of factuality-related information.[18]

On the other hand, FactBank contributes specific factuality values to the events in the TimeML-annotated corpora. The new annotation is kept in separate documents, but it is linked to the original data through the events IDs (eid attributes in TimeML EVENT tags), which are the same in both annotation layers. The FactBank annotation for that same sentence above is as shown:

(46)

Event (ID):	**Source (ID):**	**Fact. value:**
said (e22)	*author* (s_0)	CT+
infatuated (e23)	*reports_author* (s_2_s_0)	CT+
	author (s_0)	Uu
trying (e24)	*reports_author* (s_2_s_0)	PS+
	author (s_0)	Uu
impress (e25)	*reports_author* (s_2_s_0)	Uu
	author (s_0)	Uu
killing (e26)	*reports_author* (s_2_s_0)	Uu
	author (s_0)	Uu

Each event is associated with one or more pairs of relevant sources and factuality values, which express the factuality degree assigned to that event by the given source.[19] Combining the factuality values in FactBank with the structural information in TimeML-annotated corpora is of great value for developing and informing tools aimed at automatically identifying the factuality values of events.

[18] It must be pointed out, however, that none of the aforementioned issues are problems from a TimeML perspective, since its goal is not to provide a full-fledged annotation of factuality. Moreover, TimeML has been intentionally conceived of as a surface-based markup, which explains why, for instance, modal auxiliaries are recorded but not interpreted.

[19] For the sake of clarity, the example above provides both the form and the ID for events and sources, but the original FactBank annotation records only the IDs.

5.4 Annotation schema

Currently, FactBank annotation is carried out by means of stand-off markup, but it can be expressed using XML tags along the same lines as the standard version of TimeML. The annotation schema is as follows:

```
<EVENT>
      attributes  ::=   eid eiid
      eid ::= ID
      {eid ::= EventID
      EventID ::= e<integer>}
      eiid ::= ID
      {eiid ::= EventInstanceID
      EventInstanceID ::= ei<integer>}
<SOURCE_STRING>
      attributes ::= ssid
      ssid ::= ID
      {ssid ::= SourceStringID
      SourceStringID ::= s<integer>}
<RELEVANT_SOURCE/>              # a non-consuming tag
      attributes ::= rsid
      rsid ::= ID
      {rsid ::= RelevantSourceID
      RelevantSourceID ::= s<integer>[_s<integer>]*}
<FACT_VALUE/>              # a non-consuming tag
      attributes ::= fvid eiid rsid value
      fvid ::= ID
      {fvid ::= FactValueID
      FactValueID ::= f<integer>}
      eiid ::= ID
      {eiid ::= EventInstanceID
      EventInstanceID ::= ei<integer>}
      rsid ::= ID
      {rsid ::= RelevantSourceID
      RelevantSourceID ::= s<integer>[_s<integer>]*}
      value ::= 'CT+'|'PR+'|'PS+'|'CT−'|'PR−'|'PS−'|'CTu'|'Uu'
```

The EVENT tag annotates the event-denoting expressions identified as such in TimeBank, assigning them the same event and event instance IDs (attributes eid and eiid, respectively) as in that corpus. Hence, the data in FactBank is anchored to that in TimeBank by means of this tag.

The SOURCE_STRING tag marks those text strings expressing sources of factuality evaluations. In sentence (44), for instance, it will be wrapping *reports*.[20] It

[20] We follow here the same approach as TimeML of annotating only heads.

contains only one attribute, its ID (ssid), which corresponds to the sentence position of the string—starting the counting at position 1. Hence, string *reports* receives s2:

(47) Newspaper <SOURCE_STRING ssid="s2">reports</SOURCE_STRING > ...

The RELEVANT_SOURCE tag, on the other hand, is non-consuming. It presents the actual sources that are relevant for the event, corresponding to either the text author or any nested source generated by the introduction of a new source in discourse. By convention, we identify the author source as s0. And as for nested sources, their ID will consist of the ID of their textual string (i.e., the value of attribute ssid in the tag SOURCE_STRING representing them) prefixed to the IDs of the relevant sources under which they are nested (i.e., the value of attribute rsid in the tag RELEVANT_SOURCE of the nesting sources). For instance, the ID for the source *reports* presented above is s2_s0, and hence, the two relevant sources for the event *infatuated* in the previous example are:

(48)
```
<RELEVANT_SOURCE rsid="s2_s0"/>    # Referring to source 'reports'
<RELEVANT_SOURCE rsid="s0"/>       # Referring to text author
```

Finally, the FACT_VALUE tag represents the factuality value assigned by a relevant source to a given event. Because each event can have more than one factuality value assigned (as many as it has relevant sources), FACT_VALUE must be a non-consuming tag. The factuality values assigned to the event *infatuated* (e23) are annotated as:

(49)
```
<FACT_VALUE fvid="f2" eiid="ei23" rsid="s2_s0" value="CT+"/>
<FACT_VALUE fvid="f3" eiid="ei23" rsid="s0" value="Uu"/>
```

6 Annotation effort

6.1 Annotation design

The following constraints were considered in annotating the factuality of events in order to make the task more feasible, while at the same time respecting the information provided by the data.

Textual-based annotation. Judging the factuality status of an event can be influenced by what annotators know about the world. If this knowledge is allowed in the annotation, there is the risk of obtaining different judgments because of the difference in the degree of information each annotator has. Thus, the annotation must be textual-based, that is, reflecting only what is expressed in the text and avoiding any judgment based on individual knowledge.

Textual unit of information. A second issue concerns the textual unit over which annotators base their judgments. The factuality of events is mainly expressed within the containing sentence, but it is not uncommon that it is also characterized over several sentences. An event may be first presented in a totally uncommitted way (for example, embedded under a predicate expressing promise), and later on be characterized as a fact. We assume that changes like this can be handled with an adequate model of discourse structure. For the present research, we decided to constrain our annotation to information only present at the sentence level.

Complexity of the linguistic data. As we saw above, determining event factuality involves several layers of linguistic information. First, it involves the interaction of local and non-local data. Second, it may engage several relevant sources for each event, which in addition bear a nesting relation between them. Hence, if not structured adequately, the annotation task may become too complex and result in questionable markup of the data. Annotating event factuality needs to be addressed by steps that both help annotators to mentally structure and comprehend the different information layers involved, as well as allow us to partially automate certain parts of the annotation process. We divide the annotation effort into three consecutive tasks, presented in the next section.

6.2 Annotation tasks

6.2.1 Task 1: identifying source-introducing predicates (SIPs)

Given a text with the events already recognized, the annotator identified those that correspond to source-introducing predicates. SIPs were briefly described in Sect. 2.4.3 as including predicates of reporting, knowledge, and opinion, among others. They are the linguistic elements that contribute a new source to the discourse. Such new sources, which must be nested relative to any previous relevant source, play a role in assessing the factuality of the SIP event complement—*cf.* also Sect. 3.3.

This initial task allowed annotators to become familiar with both the notion of source and the notion of SIP as a marker of factuality information. Moreover, for processing purposes, Saurí (2008) shows that identifying SIPs is fundamental for automatically computing relevant sources. Hence, a corpus annotated with this kind of data is of great value for informing tools devoted to identifying opinion sources, in line with other work in the field, e.g., Bethard et al. (2004) and Choi et al. (2005).

6.2.2 Task 2: identifying sources

Sources introduced by SIPs tend to be expressed by their grammatical subject (50a), but an oblique, possibly optional, complement can also be used (50b–c). Nominal SIPs introduce sources as well (50d). In the examples below, new sources are in bold face, the SIPs underlined, and the SIP-embedded events in italics and with subindex *e*.

(50) a. In mid-2001, **Colin Powell** and **Condoleezza Rice** both publically denied that Iraq *had*$_e$ weapons of mass destruction.
b. It seemed to **him** that a girl's story about her goat *was*$_e$ more important.
c. He was told by **Cheney** that Bush had *approved*$_e$ a plan in which Libby would brief a specific New York Times reporter.
d. **Unisys Corp**.'s announcement Friday of a $648.2 million *loss*$_e$ for the third quarter showed that the company is moving even faster than expected.

The annotator was provided with text where the following information was already annotated: (a) all the SIPs in the text—obtained from the previous task, after having passed through a step validating the annotation; and (b) for each of these

SIPs, a set of elements that can potentially express the new source it introduces; that is, a set of new source candidates. New source candidates had been automatically identified by selecting NP heads holding any of the grammatical relations in the following list.[21] (In the examples, the new source candidate is marked in bold face and the SIP is underlined.)

1. Subject of any verbal predicate in the sentence.
2. Agent of a SIP in a passive construction (e.g., *The crime was reported* by the ***neighbor***.)
3. Direct object of a SIP that has, as one of its arguments, a control clause headed by another SIP (e.g., *He criticized* ***Ed*** *for saying*…).
4. Complement of preposition *to* at the beginning of a sentence (e.g., *To* ***me***, *she*…).
5. Complement of preposition *to* that is in a dependency relation with a SIP (e.g., *according to* **me**, *it seems to* **me**.)
6. Complement of preposition *of* that is in a dependency relation with a noun SIP (*the announcement of* ***Unisys*** *Corp*.).
7. Possessor in a genitive construction whose noun head is a SIP (e.g., ***Unisys*** Corp.'s *announcement*).

The annotation tool is presented in Fig. 3, where source candidates are displayed in bold face, and SIPs are in bold face and underlined.

For every SIP, the annotator selected the new source it introduces among those in the candidate set. Two exceptional situations were also accounted for:

(i) The new source did not correspond to any of the candidates in the list. The annotator would in these cases select option OTHER, and a later adjudication process would pick the adequate text item;

(ii) There was no explicit segment in the text referring to the new source—for instance, in the case of generic sources (e.g., *it was expected/assumed that*…). The annotator would then select for option NONE. The new source is interpreted as generic, and will be represented as GEN in the resulting chain expressing the relevant source (e.g., GEN_*author*).

6.2.3 Task 3: assigning factuality values

This final task was devoted to selecting the factuality value assigned to events by each relevant source. The annotators were provided with a text where every event expression was paired with its relevant sources. Hence, sentences containing events with more than one relevant source were repeated several times, each presenting a different event-relevant source pair.

The set of relevant sources for each event had been automatically computed given the new sources manually identified in the previous task, and based on the algorithm for finding relevant source chains presented in Saurí (2008). The

[21] These syntactic functions were obtained from parsing the corpus with the Stanford Parser (de Marneffe et al. 2006b).

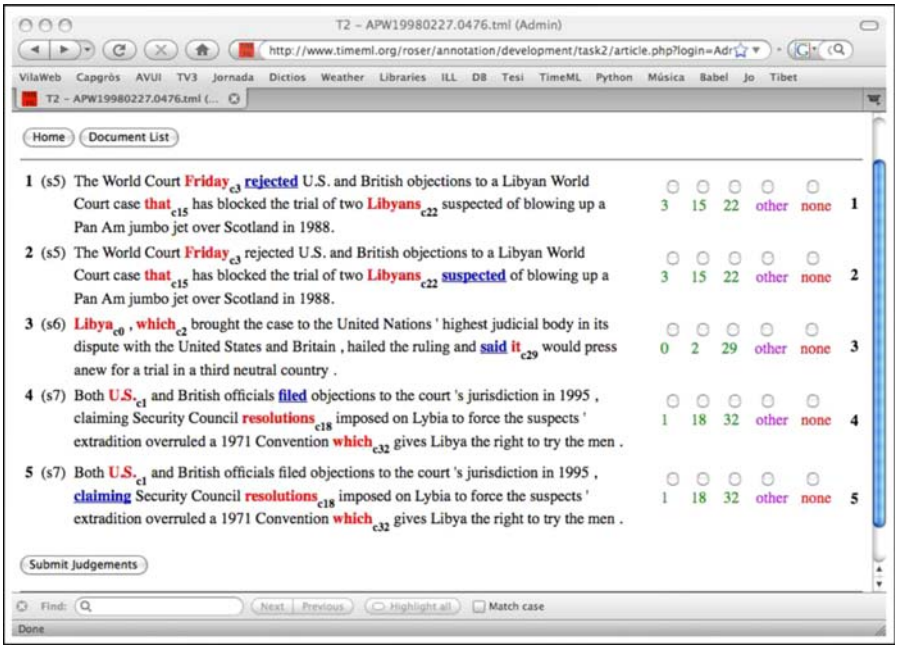

Fig. 3 Task 2 annotation screen

annotation tool (Fig. 4) displays the sentences (3rd column) with the event to be assessed in bold face and underlined. The relevant sources are in the 4th column, while the 5th column contains the factuality values to select from.

The annotator had to choose among the set of factuality values presented in Table 5, which corresponds *grosso modo* to Table 2 with the addition of values PRU and PSU. In establishing the former table, these two values were estimated as non relevant, but we wanted to confirm that at the light of real data.

Two further values were allowed as well in order to pinpoint potential limitations in our value set: OTHER, covering situations where a different value would be required (e.g., the combinations U+ and U−), or when the annotator did not know what value to select; and NA (non-applicable), for events whose factuality cannot be evaluated.

To discern among the different factuality values, the annotators were asked to apply the discriminatory tests presented in Sect. 5.1.2.

7 Results and evaluation

FactBank was annotated by a pair of annotators, both of whom were undergraduates competent in linguistics, and adjudicated by the authors. The annotator training was minimal. For each task, they read the annotation guidelines and annotated two documents, which were reviewed by us. We then met with each of them and

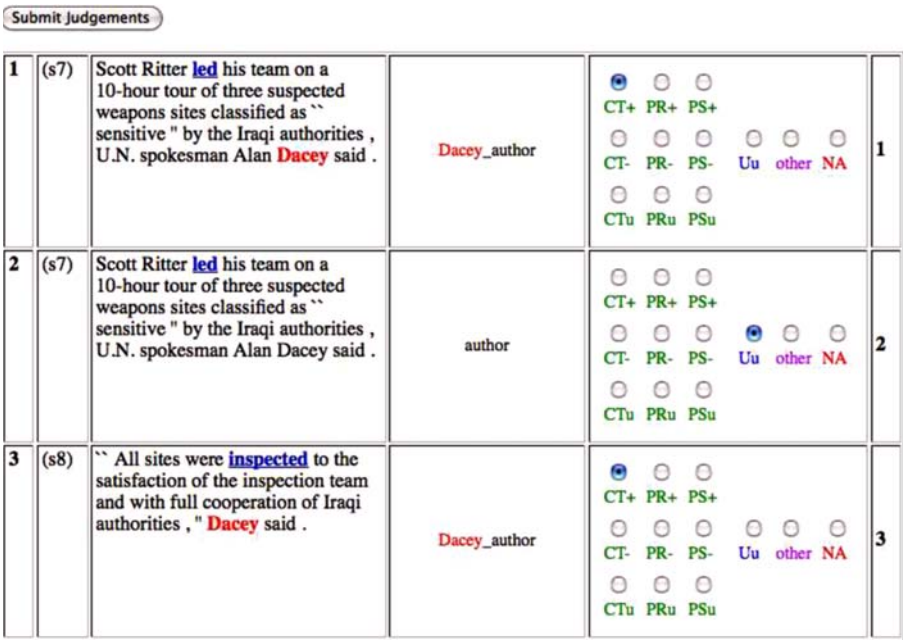

Fig. 4 Task 3 annotation screen

Table 5 Factuality values and their use

VALUE	USE
Committed values:	
CT+	According to the source, it is **certainly** the case that X
PR+	According to the source, it is **probably** the case that X
PS+	According to the source, it is **possibly** the case that X
CT−	According to the source, it is **certainly not** the case that X
PR−	According to the source it is **probably not** the case that X
PS−	According to the source it is **possibly not** the case that X
(Partially) Uncommitted values:	
CTu	The source knows whether it is the case that X or that not X
PRu	The source knows whether it is probably the case that X or that not X
PSu	The source knows whether it is possibly the case that X or that not X
Uu	The source does not know what is the factual status of the event, or does not commit to it
Other values:	
Other	Covering the following two situations: A different value is required here (e.g., U+, U−) The annotator does not know what value to assign
NA	The factuality nature of the eventuality cannot be evaluated

discussed any misunderstanding. These meetings never lasted longer than one hour. Neither of the annotators was involved in the development of the coding schema, a fact which suggests that, if the schema is adopted by other research groups, the annotation will most likely achieve a level of quality comparable to what will be reported next.

Interannotation agreement has been assessed using the *kappa* coefficient. *Kappa* is defined as $\kappa = \frac{P(A)-P(E)}{1-P(E)}$, where $P(A)$ is the observed agreement among annotators and $P(E)$ the expected agreement. The literature offers different formulas for computing $P(E)$, although the most typically used is that proposed in Cohen (1960). As pointed out in Di Eugenio and Glass (2004), however, Cohen's *kappa* suffers from skewed distributions of categories (the so-called prevalence problem), as well as from the degree to which the annotators disagree (the bias problem). In order to overcome these tendencies, they propose to report interannotation agreement by means of three values, namely, Cohen's *kappa* and two adjustments of it: one in which it is adjusted for bias, which corresponds to *kappa* coefficient as defined in Siegel and Castellan (1988), and another in which it is adjusted for prevalence, which is equal to $2P(A)-1$. The interannotation agreement for the three tasks in FactBank are reported in this way—with Cohen's *kappa* emphasized in bold face.

7.1 Evaluating task 1: identifying source-introducing predicates

The interannotation agreement achieved is $\kappa_{cohen} = \mathbf{0.88}$ over 40% of the corpus (on the number of events). The ratios obtained from the two adjusting measures are: $\kappa_{s\&c} = 0.88$ and $2P(A)-1 = 0.92$. Some of the most common cases of disagreement concern:

- SIP candidates presenting a generic source –which thus was not textually expressed (e.g., *He's **expected** to meet with Iraqi deputy prime minister later this afternoon*).
- SIP candidates which do not have an explicit event complement (e.g., *The executives didn't **disclose** the size of the expected gain*).
- SIP candidates whose event complement is expressed by an uncommon construction –i.e., not a direct object or a complement clause (e.g., *Telerate has **criticized** Dow Jones [for not disclosing those data]*).
- SIP candidates whose event complement is dislocated from its standard position, appearing either somewhere else in the sentence, or in a preceding sentence (e.g., *Har-Shefi said [she heard Amir talk about killing Rabin] but did not **tell** the police*).
- SIP candidates which appear negated. Some of the examples already provided illustrate this as well (e.g., *didn't disclose, did not tell*).
- SIP candidates which are polysemous between a SIP and an non-SIP interpretation (e.g., *It **appeared** to him that …* versus *Bunchhay **appeared** confident*, respectively).
- Speech act predicates which nevertheless do not behave as SIPs (e.g., *"It looks as though they panicked," a detective*, ***speaking** on condition of anonymity, said*).

7.2 Evaluating task 2: identifying sources

The interannotation agreement achieved for this task is: κ_{cohen} = **0.95**, $\kappa_{s\&c}$ = 0.95, and $2P(A)-1 = 0.97$, over 40% of the corpus on the number of events. Such good results come as no surprise since this is a very well-defined task, both in syntactic and semantic terms. Essentially, it requires identifying the SIP logical subject. Below are examples of common cases of disagreement. They present the SIP in bold face and the new source to be selected in bold face and underlined. If an additional expression enters consideration as a new source candidate as well, it will only be underlined. The most common cases of disagreement are those in which:

- There is a second expression in the text correfering with the new source. For example, the 1st person pronoun in a quoted fragment (e.g., "*We are going to maintain our forces in the region for a while,"* **said** *spokesman Kenneth* **Bacon**.) Another common situation was given with relative clauses (e.g., *British police* **officers** *who had been searching for Howes* **concluded** *that* …).
- The new source introduced by the SIP refers to a non-human entity (e.g., **Reports** *attributed to the Japanese foreign ministry* ***said*** …).
- The SIP is passivized and has an implicit source. For example, the entity *Unisys* in the following sentence was wrongly understood as the source introduced by *expected*: *Unisys is* **expected** *to do little better than break even in the fourth quarter.*
- The new source (generally, of a noun event) is expressed by means of a PP complement of optional nature (e.g. *Netanyahu*'s *comments last week were in response to* **signals** *from* **Syria** *that* it *wants to renew the long-stalled peace talks.*)

7.3 Evaluating task 3: assigning factuality values

Interannotion agreement for this task scores at κ_{cohen} = **0.81**, $\kappa_{s\&c}$ = 0.81, and $2P(A)-1 = 0.80$, over the 30% of the corpus (in terms of number of events). We consider this a very acceptable result given the complexity of the task. Refer to Sect. 7.5 for further discussion of these numbers.

An analysis of the disagreement instances from 10% of the corpus showed that around two thirds of them (65%) are cases of true ambiguity, originating from different kinds of constructions. Some of the most recurring ones concern:

- **The scope of a reporting predicate**, or in other words, the span of the attributed fragment. In (51), for example, the reporting predicate (in bold face) can be interpreted as scoping over both events *want* and *traveled*, or only over *traveled*. In the second interpretation, *want* is characterized as a fact (CT+) according to the author, since she is the one directly asserting it. On the other hand, if *want* is reported by the investigator (first interpretation), the author is uncommitted (Uu).

  (51) Authorities want to question the unidentified woman who allegedly traveled with Kopp, **according** to an investigator quoted by the newspaper.

- **Syntactic constructions typically triggering a presupposition** (e.g., relative clauses, temporal clauses, appositions), when embedded under a reporting predicate (52). Annotators disagree on whether the presupposition is projected to the main clause. In our terms, this disagreement can be stated as to whether the author of the text assesses the embedded event (underlined below) as a fact (CT+) or remains uncommitted (Uu).

  (52) The killing of Dr. Barnett Slepian, a gynecologist in Buffalo who performed abortions, has become a factor in at least two campaigns in New York, **say** political consultants.

  This is a highly common case of disagreement among annotators, but interestingly, there is no clear position about this in the linguistic literature. See, e.g., Karttunen (1973), Geurts (1998), or Glanzberg (2003).
- **Event-denoting nouns**, especially when embedded under a reporting predicate. Similarly to the previous case, there is disagreement concerning whether the nominal event (underlined below) is understood as a fact not only by the subject of the reporting event, but by the text author as well.

  (53) Last week, FBI Director Louis Freeh, on an official visit to Mexico, asked Mexican authorities to join the hunt for Kopp, federal officials have **said**.

  Noun events are a highly frequent source of disagreement, which comes as no surprise: the annotation guidelines did not address at all how to deal with them. The omission was deliberate, however, given the complexity concerning noun interpretation, which touches on thorny issues such as definiteness, or involves ontological-based information. For example, the event complement of *block* will receive a factual or non-factual interpretation depending on the word sense assumed for *block*. Compare, e.g., *block the offer*, where *block* = *refuse* (factual) with *block the trial*, where *block* = *prevent* (non-factual) (Pustejovsky and Rumshisky 2008).
- **Present perfect clauses**. They allow for both a factual (CT+) or an underspecified (Uu) interpretation of the mentioned event:

  (54) **Extending** membership to these three democracies helps to stabilize a region that historically has been the staging ground for many of the disasters of this century.

- **Purpose clauses**. Their main event tends to be characterized as underspecified (Uu), but in some contexts it can be understood as factual as well.

  (55) The murder of Dr. Barnett Slepian is the latest depraved assault on the small number of courageous physicians who still risk their safety to **perform** legal abortions.

- **ESPs** that can be interpreted as qualifying the factuality of their embedded events in different ways. Some of the frequent disagreements concerned predicates *believe*, *think*, *admit*, *agree*, *decide*, and *help*. We comment on a couple of them below:

Believe: One annotator tended to interpret that the cognizer participant (in bold face below) characterizes the embedded events (underlined) as probable (PR+), whereas the other opted for a factual position (CT+).

(56) The hair was in a packet that was found buried in the ground more than 50 feet from a tree where **police** *believe* the sniper stood and fired.

Admit: It tends to have a factive interpretation, which in our terms means that both the author of the text and the cognizer commit to its embedded event (underlined) as a fact (CT+). However, there were several sentences in which one of the annotators interpreted *admit* as a regular reporting predicate –i.e., a predicate that presents the cognizer committing to the embedded event as a fact (CT+) and the text author as uncommitted (Uu).

(57) **Police** *admit* that despite a worldwide search and $1 million reward, they don't know where he is.

– **Modal auxiliaries**, **such as *can* and *would***, which are polysemous by nature. The event in bold face in the sentence below, for example, was characterized as both possible (PS+) and uncertain (Uu).

(58) With their membership secure, they also can **participate** in planning for the NATO summit in April.

The remaining cases of disagreement are true errors. The most common of them concern: (a) wrong (inaccurate or careless) application of the guidelines—especially in cases of negated events, conditional constructions, and sentences presented in quotation marks; and (b) misunderstanding a nested source, mainly those involving generic sources (e.g., *it is expected/said/*…).

7.4 Data distribution

The distribution of data in the third annotation task provides some insight about the adequacy of the set of factuality values proposed in Sect. 5.1. Table 6 shows the distribution of each value in the AQUAINT TimeML Corpus, corresponding to 16.4% of the overall number of events.

As is apparent from the table, the data is quite skewed, favoring two main categories: CT+ and Uu. The prevalence of the first value should come as no surprise. We are analyzing data from news reports, which are about things that have

Table 6 Distribution of factuality values, after adjudication (AQUAINT TimeML Corpus)

Value	CT+	CT−	Ctu	PR+	PR−	Pru	PS+	PS−	Psu	Uu	Other	NA
Counts	1276	51	1	25	0	0	33	2	0	804	0	5
%	58.1	2.3	0.05	1.15	0	0	1.5	0.1	0	36.6	0	0.2

Table 7 Distribution of factuality values in different contexts (AQUAINT TimeML Corpus)

Value	CT+	CT−	Ctu	PR+	PR−	Pru	PS+	PS−	Psu	Uu	Other	NA
#Simple	794	31	0	2	0	0	4	0	0	156	0	0
#Embed	482	20	1	23	0	0	29	2	0	648	0	5
%Simple	36.1	1.4	0	0.1	0	0	0.2	0	0	7.1	0	0
%Embed	22	0.9	0.05	1.05	0	0	1.3	0.1	0	29.5	0	0.2

happened in the world. The high frequency of Uu, however, is less obvious, especially considering that categories with a modality degree lower than CT (i.e., PR and PS) or a non-positive polarity (CT−, CTu, PR−, etc.) appear to be very infrequent in the analyzed corpus, occupying only the 5.1% of the events. But a closer look at the data explains the high frequency of Uu.

Table 7 shows the number and percentage of each value for the same events above, split into two subcorpora: one containing those events embedded under a SIP (such as *say, know, believe*), and the other containing events placed either at the main level of the sentence, or embedded by predicates other than SIPs (like *offer, want, improve*, or *manage*)—refer to Sect. 2.4.3 for the distinction between SIPs and NSIPs. We call them events in *embedded* and *simple* contexts, respectively, although the second group contains events that are syntactically embedded as well.

As can be seen, the frequency of value Uu is remarkably larger in embedded contexts than in simple ones. In fact, almost one third of the factuality values in embedded contexts correspond to instances of Uu (29.5%). The reason for this is clear. Events embedded by SIPs have at least two sources: the text author and the participant expressed by the logical subject of the SIP (e.g., *John* in *John said/thought that Mary came*.). As it happens, most of these embedded contexts express the subject participant's opinion (e.g., *that Mary came*), with regards to which the author remains uncommitted (Uu).

On the other hand, many of the 7.1% of events in simple contexts receiving the value Uu are subordinated by predicates of volition (e.g., *want*), commitment (*offer*), or command (*order*), among others. They are therefore events of underspecified factuality.

It is worthwhile noting that there is a qualitative difference between the meaning of value Uu when used in simple contexts and when used to express non-commitment of a source, like in contexts of reporting. The uncertain nature of the events in the first situation can be changed or overwritten if more information is provided about that event, whereas the uncommitted values in the second situation can be changed into a more informative value if a measure of source reliability is applied.

A second observation in light of the annotated data concerns the values PRu and PSu. They were not considered applicable when establishing the set of factuality values in Sect. 5.1, but were nevertheless presented to the annotators in order to test their utility. A look at the contingency matrix for task 3 (Table 8) shows that there was no event judged as either PRu or PSu in 30% of the corpus, which confirms the hypothesis previously assumed.[22]

[22] As a matter of fact, there was no event judged as such throughout the whole corpus.

Table 8 Contingency matrix for task 3 (over 30% of the corpus)

	CT+	CT−	Ctu	PR+	PR-	Pru	PS+	PS−	Psu	Uu	Other	NA	Total
CT+	2483	1	0	21	0	0	2	0	0	97	1	0	2605
CT−	17	136	0	0	1	0	0	0	0	15	0	0	169
CTu	1	0	0	0	0	0	0	0	0	2	0	0	3
PR+	5	0	0	38	0	0	0	0	0	8	0	1	52
PR−	1	0	0	0	4	0	0	0	0	2	0	0	7
PRu	0	0	0	0	0	0	0	0	0	0	0	0	0
PS+	1	0	0	1	0	0	34	0	0	25	0	0	61
PS−	0	0	0	0	0	0	0	1	0	1	0	0	2
PSu	0	0	0	0	0	0	0	0	0	0	0	0	0
Uu	189	21	0	31	6	0	23	0	0	1615	2	6	1893
Other	2	0	0	1	0	0	0	0	0	0	0	0	3
NA	6	0	0	0	0	0	0	0	0	0	0	0	6
Total	2705	158	0	92	11	0	59	1	0	1765	3	7	4801

The data also validates the need for distinguishing between two uncertain modalities, probable (PR) and possible (PS). Table 8 demonstrates that there is barely confusion between the two categories between annotators.

An interesting observation concerns the value OTHER, which does not reflect any linguistic distinction but was introduced in the set to provide annotators with an alternative in cases of ambiguity (more than one value could be chosen) or doubt (they were not sure what to choose). As it turned out, OTHER was selected by each annotator in only 3 out of 4801 instances of pairs event-source in 30% of the corpus (Table 8). This suggests that annotators tend to look for a non-ambiguous factuality interpretation of the event, very much in the same way speakers do when using language in natural contexts.

7.5 Interpreting the results

The interannotation agreement obtained for characterizing events with respect to their degree of factuality ($\kappa_{cohen} = \kappa_{s\&c} = 0.81$, Task 3) is both surprising and extremely satisfying considering the degree of difficulty of the task. The significance of this result can be better appreciated when compared with that attained in the (so far only) comparable experiment on annotating certainty degrees (Rubin 2007). She obtained $\kappa_{cohen} = 0.15$, which improved to 0.41 when stricter annotation instructions were provided.[23] Hence, the agreement achieved in FactBank demonstrates that the annotation framework established for event factuality (the set of factuality values together with a battery of tests) is rich enough for expressing the necessary

[23] Rubin's approach and ours are not completely equivalent, since she annotates only sentences where there are "explicit markers of certainty", whereas we assume that factuality is a value affecting all events in text. In addition, her system does not consider polarity as part of the information to identify.

distinctions, but at the same time is not too complex for annotators to agree in judgment in a significant way.

The adequacy of our annotation specification is further attested when comparing our results with those achieved for annotating linguistic structures better constrained by syntactic and lexical means, and for which there already exists a fairly well established descriptive model in the field; e.g., semantic role labeling or discourse structure. For example, the agreement of identifying semantic roles of nominal predicates is about 85% for argument roles, lower for adjuncts (Meyers et al. 2004); identifying both argument and adjunct roles for verbs leads to $\kappa_{s\&c} = 0.91$ (Palmer et al. 2005). And the agreement of identifying argument extents in discourse relations is 85.1% and 90.2% for arguments of implicit and explicit connectors, respectively (Prasad et al. 2008). Note in addition that these annotation tasks do not entail the same degree of unresolved ambiguity that we observed for the factuality value of events; *cf.* Sect. 7.3.

As it can be appreciated, the difference between our score and those achieved in the tasks mentioned above is placed within a range of 0.1 points (or 10%, in percentile terms), which proves that the annotation specification proposed here is the appropriate direction to take for identifying and representing the factuality degrees of event mentions in discourse.

8 Conclusions

Event factuality is an important component for representing events in discourse, but identifying it presents a three-fold challenge. First, factuality in itself is a continuous system. Hence, in order to obtain consistently annotated data, it is important to establish a set of discrete factuality values which are grounded both on linguistic intuitions and on commonsense reasoning, so that they reflect the distinctions that speakers naturally apply. Second, factuality is in many cases the result of different interacting markers. They can all be in the local context of the event, but it is also common for them to occur at different levels. And finally, the factuality value assigned to events in text must be relative to the relevant sources at play, which may be one or many.

In this article, we introduced FactBank, a corpus of events annotated with factuality information. We first defined event factuality, identifying its markers and establishing the descriptive framework to be used for annotation, and then explained the annotation effort in detail. FactBank contributes a semantic layer of factuality information on top of the grammar-based layer provided in TimeBank. When combined, these two layers are of great value for developing tools aimed at automatically identifying the factuality values of events. FactBank will be made available to the community through the Linguistic Data Consortium (LDC).

The interannotation agreement scores obtained are very positive. Specifically, for the task of selecting the factuality value assigned to events by each of their relevant sources, we achieved $\kappa = 0.81$ over 30% of the corpus. In addition to validating the overall quality of FactBank, such a positive result suggests that event factuality as modeled in our work is well-grounded, and that its identification is achievable using

an approach along the lines proposed here. Furthermore, the interannotation agreement score also provides some hints about the results we can expect when attempting to identify factuality by automatic means, as it has been initially explored in Saurí (2008).

These results are the product of systematic work along three fronts. First, we identified and analyzed the linguistic mechanisms involved in event factuality, which concern all levels within the linguistic system: grammatical, lexical, syntactic, and discourse. This analysis step is important in order to have a complete understanding of the problem and what it involves (e.g., interaction among factuality markers, presence of several sources), and then be able to structure the annotation procedure adequately. Second, we created a consistent annotation framework (set of values and discriminatory tests), based on both theoretical findings, linguistic intuitions, and commonsense reasoning. Finally, we established a simple annotation procedure, divided into even simpler, sequential subtasks. This makes it possible for complex tasks to be built on top of earlier, simpler ones, and at the same time allows annotators to become incrementally familiar and more comfortable with the complexity of the problem.

Acknowledgments We are very grateful to Marc Verhagen, Toni Badia, Lauri Karttunen, Rick Alterman, Sabine Bergler, Adam Meyers, and Silvia Pareti for their valuable comments and helpful discussion regarding this research. We also want to extend thanks to four anonymous reviewers for their constructive suggestions, which helped improve the original manuscript. All errors and mistakes are responsibility of the authors. This work is been supported by a grant to Prof. Pustejovsky, NAVAIR Contract No. N61339-06-C-0140.

References

ACE (2008). *ACE (Automatic Content Extraction) English annotation guidelines for relations* (Version 6.0 – 2008.01.07 ed.). Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/ACE.

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford, England: Oxford University Press.

Andreevskaia, A., & Bergler, S. (2006). Mining WordNet for fussy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the 11th conference of the European chapter of the Association for the Computational Linguistics, EACL-2006*.

Asher, N. (1993). *Reference to abstract objects in English*. Dordrecht, The Netherlands: Kluwer Academic Press.

Bach, K., & Harnish, R. M. (1979). *Linguistic communication and speech acts*. Cambridge, Massachusetts, USA: The MIT Press.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *17th International conference on computational linguistics* (pp. 86–90).

Bergler, S. (1992). Evidential analysis of reported speech. PhD thesis, Brandeis University.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In *Proceedings of AAAI spring symposium on exploring attitude and affect in text*.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text, 9*(1), 93–124.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. v. Kuppevelt & R. W. Smith (Eds.), *Current and new directions in discourse and dialogue*. Springer.

Chafe, W. (1986). Evidentiality in English conversation and academic writing. In W. Chafe & J. Nichols (Eds.), *Evidentiality: The linguistic coding of epistemology*. Norwood, New Jersey, USA: Ablex Publishing Corporation.

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the HLT/EMNLP 2005*. Vancouver, Canada.

Clemen, G. (1997). The concept of hedging: Origins, approaches and definitions. In R. Markkanen & H. Schröder (Eds.), *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts* (pp. 235–248). Berlin; New York: Walter de Gruyter.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 10*, 37–46.

Condoravdi, C., Crouch, R., van den Berg, M., Everett, J., Stolle, R., Paiva, V., & Bobrow, D. (2001). Preventing existence. In *Proceedings of the conference on formal ontologies in information systems (FOIS)*, Ogunquit, Maine, USA.

Dave, K. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of World Wide Web conference 2003*.

de Haan, F. (1997). *The interaction of modality and negation: A typological study*. New York, USA: Garland.

de Haan, F. (2000). The relation between modality and evidentiality. In R. Müller & M. Reis (Eds.), *Modalität und Modalverben im Deutschen*. Hamburg, Germany: Helmut Buske Verlag.

de Marneffe, M.-C., MacCartney, B., Grenager, T., Cer, D., Rafferty, A., & Manning, C. D. (2006a). Learning to distinguish valid textual entailments. In *Second PASCAL RTE Challenge (RTE-2)*.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006b). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: a second look. *Computational Linguistics, 30*, 95–101.

Dor, D. (1995). Representations, attitudes and factivity evaluations. An epistemically-based analysis of lexical selection. PhD thesis, Stanford University.

Geurts, B. (1998). Presuppositions and anaphors in attitude contexts. *Linguistics and Philosophy, 21*, 545–601.

Givón, T. (1993). *English grammar. A function-based introduction*. Amsterdam, The Netherlands: John Benjamins.

Glanzberg, M. (2003). Felicity and presupposition triggers. In *University of Michigan Workshop in Philosophy and Linguistics*. Michigan, USA.

Halliday, M. A. K. (1994). *An introduction to Functional Grammar* (2nd ed.). London, England: Edward Arnold.

Halliday, M. A. K., & Matthiessen, C. M. (2004). *An introduction to Functional Grammar*. London, England: Hodder Arnold.

Hickl, A., & Bensley, J. (2007). A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the workshop on textual entailment and paraphrasing* (pp. 171–176). Prague, Czech Republic.

Hooper, J. B. (1975). On assertive predicates. In J. Kimball (Ed.), *Syntax and semantics, IV* (pp. 91–124). New York, USA: Academic Press.

Horn, L. R. (1972). On the semantic properties of logical operators in English. PhD thesis, UCLA. Distributed by the Indiana University Linguistics Club in 1976.

Horn, L. R. (1989). *A natural history of negation*. Chicago, USA: University of Chicago Press.

Huddleston, R. (1984). *Introduction to the grammar of English*. Cambridge, England: Cambridge University Press.

Karttunen, L. (1970). Implicative verbs. *Language, 47*, 340–358.

Karttunen, L. (1973). Presuppositions of compound sentences. *Linguistic Inquiry, 4*(2), 169–193.

Karttunen, L., & Zaenen, A. (2005). Veridicity. In G. Katz, J. Pustejovsky, & F. Schilder (Eds.), *Dagstuhl seminar proceedings*. Schloss Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum (IBFI).

Kiefer, F. (1987). On defining modality. *Folia Linguistica, XXI*, 67–94.

Kiparsky, P., & Kiparsky, C. (1970). Fact. In M. Bierwisch & K. E. Heidolph (Eds.), *Progress in linguistics. A collection of papers* (pp. 143–173). The Hague: Mouton.

Koenig, J.-P., & Davis, A. R. (2001). Sublexical modality and the structure of lexical semantics. *Linguistics and Philosophy, 24*, 71–124.

Kratzer, A. (1991). Modality. In A. van Stechow & D. Wunderlich (Eds.), *Semantik: Ein internationales Handbuch der zeitgenoessischen Forschung* (pp. 639–650). Berlin, Germany: Walter de Gruyter.

Light, M., Qiu, X. Y., & Srinivasan, P. (2004). The language of Bioscience: Facts, speculations, and statements in between. In *BioLINK 2004: Linking biological literature, ontologies, and databases* (pp. 17–24).

Lyons, J. (1977). *Semantics*. Cambridge, England: Cambridge University Press.

Martin, J. R., & White, P. R. R. (2005). *Language of evaluation: Appraisal in English*. Palgrave Macmillan.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. (2004). The NomBank project: An interim report. In *Proceedings of frontiers in corpus annotation workshop. HLT-NAACL*.

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004). The Penn Discourse TreeBank. In *Proceedings of LREC 2004*.

Mushin, I. (2001). *Evidentiality and epistemological stance*. Amsterdam/Philadelphia: John Benjamin.

Nairn, R., Condoravdi, C., & Karttunen, L. (2006). Computing relative polarity for textual inference. In *Inference in Computational Semantics, ICoS-5*.

Palmer, F. R. (1986). *Mood and modality*. Cambridge, England: Cambridge University Press.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71–105.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 115–124.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the EMNLP 2002*.

Polanyi, L., & Zaenen, A. (2005). Contextual lexical valence shifters. In J. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: Theories and applications*. New York, NY, USA: Springer-Verlag.

Pradhan, S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007). OntoNotes: A unified relational semantic representation. In *Proceedings of IEEE international conference on semantic computing, ICSC 2007* (pp. 517–526).

Prasad, R., Dinesh, N., Lee, A., Joshi, A., & Webber, B. (2007). Attribution and its annotation in the Penn Discourse TreeBank. *Traitement Automatique des Langues, 47*(2), 43–63.

Prasad, R., Dinesh, N., Lee, A., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*, Marrakesh, Morocco.

Pustejovsky, J., Castano, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., & Katz, G. (2003). TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5, fifth international workshop on computational semantics*.

Pustejovsky, J., Verhagen, M., Saurí, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, R., & Setzer, A. (2006). TimeBank 1.2. Linguistic Data Consortium (LDC). Philadelphia, PA. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08.

Pustejovsky, J., Knippen, B., Littman, J., & Saurí, R. (2005). Temporal and event information in natural language text. *Language Resources and Evaluation, 39*(2), 123–164.

Pustejovsky, J., & Rumshisky, A. (2008). Between chaos and structure: Interpreting lexical data through a theoretical lens. *Special Issue of International Journal of Lexicography in Memory of John Sinclair, 21*(3), 337–355.

Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1985). *A comprehensive grammar of the English language*. London, England: Longman.

Read, J., Hope, D., & Carroll, J. (2007). Annotating expressions of appraisal in English. In *Proceedings of the linguistic annotation workshop*, Prague. Association for Computational Linguistics, ACL.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th conference on natural language learning (CoNLL 2003)*.

Rubin, V. L. (2006). Identifying certainty in texts. PhD thesis, Syracuse University.

Rubin, V. 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*.

Rubin, V. L., Liddy, E. D., & Kando, N. (2005). Certainty identification in texts: Categorization model and manual tagging results. In J. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: Theories and applications*. New York, USA: Springer-Verlag.

Saurí, R. (2008). A factuality profiler for eventualities in text. PhD thesis, Brandeis University.

Saurí, R., & Pustejovsky, J. (2007). Determining modality and factuality for text entailment. In *Proceedings of the first IEEE international conference on semantic computing*, Irvine, CA, USA.

Saurí, R., & Pustejovsky, J. (2008). From structure to interpretation: A double-layered annotation for event factuality. In *Proceedings of the second linguistic annotation workshop (The LAW II). LREC 2008*, Marrakesh, Morocco.

Saurí, R., Verhagen, M., & Pustejovsky, J. (2006a). Annotating and recognizing event modality in text. In *19th International FLAIRS conference, FLAIRS 2006*. The Florida Artificial Intelligence Research Society.

Saurí, R., Verhagen, M., & Pustejovsky, J. (2006b). SlinkET: A partial modal parser for events. In *Proceedings of LREC 2006*, Genoa, Italy.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. Boston, MA, USA: McGraw Hill.

Snow, R., & Vanderwende, L. (2006). Effectively using syntax for recognizing false entailment. In *Proceedings of HLT-NAACL 2006*.

Stoyanov, V., & Cardie, C. (2008). Annotating topics of opinions. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELDA.

Tatu, M., & Moldovan, D. (2005). A semantic approach to recognizing textual entailment. In *Proceedings of HLT/EMNLP* (pp. 371–378).

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th ACL*, 417–424.

Van Valin, R. D., & LaPolla, R. J. (1997). *Syntax. Structure, meaning and function*. Cambridge, England: Cambridge University Press.

Verhagen, M., Stubbs, A., & Pustejovsky, J. (2007). Combining independent syntactic and semantic annotation schemes. In *Proceedings of the Linguistic Annotation Workshop* (pp. 109–112). ACL. Prague, Czech Republic.

Waugh, L. R. (1995). Reported speech in journalistic discourse: The relation of function and text. *Text, 15*(1), 129–173.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2), 165–210.

Wiebe, J. M. (2000). Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000)*.

Wierzbicka, A. (1987). *English speech act verbs. A semantic dictionary*. Sydney, Australia: Academic Press.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. In *Proceedings of the HLT/EMNLP 2005 Demonstration Abstracts* (pp. 34–35). Vancouver, Canada.

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004)*.

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based analysis. *Computational Linguistics, 31*(2), 249–287.