From structure to interpretation: A double-layered annotation for event factuality

Roser Saurí and James Pustejovsky

Lab for Linguistics and Computation
Computer Science Department
Brandeis University
{roser,jamesp}@cs.brandeis.edu

Abstract

Current work from different areas in the field points out the need for systems to be sensitive to the factuality nature of events mentioned in text; that is, to recognize whether events are presented as corresponding to real situations in the world, situations that have not happened, or situations of uncertain status. Event factuality is a necessary component for representing events in discourse, but for annotation purposes it poses a representational challenge because it is expressed through the interaction of a varied set of structural markers. Part of these factuality markers is already encoded in some of the existing corpora but always in a partial way; that is, missing an underlying model that is capable of representing the factuality value resulting from their interaction. In this paper, we present FactBank, a corpus of events annotated with factuality information which has been built on top of TimeBank. Together, TimeBank and FactBank offer a double-layered annotation of event factuality: where TimeBank encodes most of the basic structural elements expressing factuality information, FactBank adds a representation of the resulting factuality interpretation.

1. Introduction

In the past decade, most efforts towards corpus construction have been devoted to encoding a variety of semantic information structures. For example, much work has gone to annotating the basic units that configure propositions (PropBank, FrameNet) and the relations these hold 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]

The field is moving now towards finding platforms for unifying them in an optimal way –e.g., Pradhan et al. (2007); Verhagen et al. (2007). It therefore seems we are at a point where the first elements for text understanding can be brought together.

Nonetheless, current work from different areas in the field points out the need for systems to be sensitive to an additional level of information; namely, that conveying whether events in text are presented as corresponding to real situations in the world, situations that have not happened, or situations of uncertain status. We refer to this level as *event factuality*.

The need for this further type of information is demonstrated in highly domain-oriented disciplines such as bioinformatics (Light et al., 2004), as well as more genre-oriented tasks. For example, Karttunen & Zaenen (2005) discusses the relevance of veridicity for IE. 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. And in the scope of textual entailment, it has been taken as a basic feature in some of the systems participating in (or using the data from) previous PASCAL RTE challenges.

For example, Tatu & Moldovan (2005) treat intensional contexts, de Marneffe et al. (2006) look at features accounting for the presence of polarity, modality, and factivity markers in the textual fragments, while Snow & Vanderwende (2006) check for polarity and modality scoping over matching nodes in a graph. Most significantly, the system that obtained the best absolute result in the three RTE challenges, scoring an 80% accuracy (Hickl & 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 –either asserted, presupposed, or implicated.

Event factuality is a necessary component for representing events in discourse, together with other levels of information such as argument structure or temporal information. 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. For instance, it is basic for temporally ordering the events in a given text.

For annotation purposes, however, it poses a representational challenge. The factuality of events is expressed through the interaction of elements from different linguistic categories. It involves, for instance, polarity (events can be presented as positive or negative) as well as modality – epistemic modality, for instance, expresses the degree of certainty of a source about what is asserted, and events qualified with other types of modality are generally presented as mere possibilities. 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 somebody else) or mood (e.g., indicative vs. subjunctive). Factuality is also a component in the semantics of specific syntactic structures with presuppositional effects (e.g., appositions and relative clauses), as well as certain types of

[1] The main references for these corpora are: PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998), RST Corpus (Carlson et al., 2002), Penn Discourse TreeBank (Miltsakaki et al., 2004), GraphBank (Wolf & Gibson, 2005), TimeBank (Pustejovsky et al., 2003), MPQA Opinion Corpus (Wiebe et al., 2005).

predicates –most notoriously, the so-called factive and implicative predicates, but also others; compare, for instance, the effect that *decision* in (1a) and *refusal* in (1b) have on the factuality status of the underlined event.

(1) a. A senior Russian politician has hailed a **decision** by Uzbekistan to shut down a United States military base.

b. A senior Russian politician has hailed the **refusal** by Uzbekistan to shut down a United States military base.

Part of these factuality markers are already encoded in some of the existing corpora (for example, TimeBank annotates polarity particles, modality operators, as well as the aforementioned predicates), but always in a partial way; that is, missing an underlying model capable of representing the factuality value that results from their interaction.

In this paper, we introduce FactBank, a corpus of events annotated with factuality information which 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.

In the next section, we set the linguistic grounding of our work by defining event factuality as a semantic property of events, establishing its possible values, and identifying its structural markers. Then, section 3 presents the main challenges for automatically recognizing it, which motivate the double-layered corpus annotation. We review some of the existing corpora where this information has already been annotated in section 4. Finally, section 5 focuses on FactBank, which is evaluated in section 6.

2. Linguistic foundations

2.1. What is event factuality

Eventualities in discourse can be couched in terms of a veridicality axis that ranges from truly factual to counterfactual, passing through a whole spectrum of degrees of modality. In some contexts, the factual status of events is presented with absolute certainty. Events are then characterized as *facts* (2) or *counterfacts* (5). Other contexts introduce different shades of uncertainty. Depending on the polarity, events are then qualified as *possibly factual* (3) or *possibly counterfactual* (4).

(2) Five U.N. inspection teams visited a total of nine other sites.

(3) United States may extend its naval quarantine to Jordan's Red Sea port of Aqaba.

(4) They may not have enthused him for their particular brand of political idealism.

(5) The size of the contingent was not disclosed.

Factuality can therefore be characterized as involving polarity and modality (more precisely, epistemic modality). Polarity is a discrete category with two values, positive and negative. Epistemic modality expresses the speaker's degree of commitment to the truth of the proposition (Palmer, 1986), which ranges from uncertain (or possible) to absolutely certain (or necessary). For methodological reasons, however, we need a discrete categorization of that system.

2.2. Factuality values

Within modal logic, two operators are typically used to express a modal context: necessity ($\Box$) and possibility ($\Diamond$); e.g., Lewis (1968). 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 & 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. He presents modality as a continuous category. Yet, he provides a good grounding for differentiating the three major modality degrees just mentioned. Based on that, we represent factuality by means of the features in Table 1:

Table 1: Factuality values

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

The factual value of events is then presented as a tuple $<mod, pol>$, containing a modality and a polarity value.[2] The polarity axis divides into positive, negative, and unknown, while the modality axis distinguishes among certain (CT), probable (PR), possible (PS), and unknown (UN). The *unknown* values are added to account for cases of uncommitment.

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 partially underspecified value, <CT,u>, is for cases where there is total certainty about the factual nature of the event but it is not clear, however, what the output is –e.g., (6). The fully underspecified <U,u>, on the other hand, is used when any of the following situations applies: (*i*) The source does not know what is the factual status of the event, as in (7a); (*ii*) the source is not aware of the possibility of the event –e.g., (7b); or (*iii*) the source does not overtly commit to it –e.g., (7c). The following examples illustrate each of these preceding situations for the underlined event when evaluated by source *John*:

(6) **John** knows whether Mary came.

(7) 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.

For simplicity, in what follows the factuality values will be represented in the abbreviated form of CT+, PR−, Uu, etc.

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

2.3. Discriminatory tests

In characterizing the factuality of events, the polarity parameter offers no problem, but distinguishing between the modality values (e.g., between *possible* and *probable*) is not always evident. In order to determine the modality parameter, we designed a battery of tests based on the logical relations considered in Horn (1989) to pinpoint the basic categories of epistemic modality; i.e., Law of Contradiction and Law of Excluded Middle. They are copredication tests.

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 (8) can be continued by either fragment in (10), 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 sentence (9), where the underlined event is explicitly characterized as probable.

(8) Iraq has agreed to allow Soviets in Kuwait to leave.

(9) Soviets in Kuwait will most probably leave.

(10) a. ... They will take the plane tomorrow early in the morning. (CT+)
b. ... However, most of them decided to remain there. (CT−)

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*. In the examples below, the symbol # is used to express that there is some sort of semantic anomaly.

(11) a. Hotels are only thirty (CT+) percent full.
b. #... but it is possible that they aren't (PS−).

(12) 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:

(13) a. I *think* it's not going to change (PR−) for a couple of years.
b. ... but it *could* happen otherwise. (PS+)

(14) a. He *probably* died (PR+) within weeks or months of his capture.
b. ... but it is also possible that the kidnappers kept him alive for a while. (PS-)

In (13), 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 (13b) can be added without creating any semantic anomaly. A similar situation is presented in (14): the adverb *probably* is characterizing the event as PR+, and the additional fragment presents the possibility of things being otherwise.

Probable (PR) versus possible (PS):

As seen, both degrees of uncertainty (PR and PS) accept copredication with PS in a context of opposite polarity. However, only the lowest degree of uncertainty (PS) accepts copredication with PR in a context of opposite polarity.

(15) a. I *think* it's not going to change (PR−) for a couple of years.
b. #... but it *probably* will. (PR+)

(16) a. It *may* not change (PS−) for a couple of years.
b. ... but it most *probably* will. (PR+)

Table 2 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, abbreviated as $\text{EM}_{subindex}$. EM expresses the epistemic modality value of the context to be copredicated to the original sentence, whereas $subindex$ indicates its polarity: $=$ means context of the same polarity, and op, context of opposite polarity.

Table 2: Tests for discriminating among modality degrees.

	CT$_=$	**CT***op*	**PR***op*	**PS***op*
U	ok	ok	ok	ok
PS	ok	#	ok	ok
PR	ok	#	#	ok
CT	ok	#	#	#

For example, given an event e presented under a context of negative polarity in its original sentence, test PR_{op} requires the creation of a new fragment in which e is used in a context where the modality degree is *probable* and the polarity is positive: PR+.[3]

(17) Original: I *think* it's not going to change. (PR−)
Testing e_2 with PR_{op}: #... but it probably will. (PR+)

2.4. Factuality markers

Event factuality in natural language is marked by both lexical items and syntactic constructions.

2.4.1. Lexical Markers

Event Selecting Predicates (ESPs). These are predicates (verbs, nouns, or adjectives) that select for an argument denoting an eventuality of any sort. Syntactically, they subcategorize for *that*-, gerundive-, and *to*- clauses, or NPs headed by event-denoting nouns. The ESPs in (18) are in bold face; their embedded events, underlined.

(18) a. Uri Lubrani also **suggested** Israel was **willing** to withdraw from southern Lebanon.
b. Kidnappers **kept** their **promise** to kill a store owner they took hostage.

[3] As appreciated, test CT= is non-discriminative. It is added there because, when combined with CP*op*, it allows to identify U values from the rest.

ESPs contribute to characterizing the factuality of the event denoted by its complement. For example, complements to weak assertive predicates (Hooper, 1975) (*think*, *suppose*) are depicted as not totally certain; complements of reporting predicates (Bergler, 1992) are presented as certain according to a particular source; factive (*regret, know*) and implicative predicates (*manage, prevent*) characterize their embedded complements as either factual or counterfactual (Kiparsky & Kiparsky, 1970; Karttunen, 1970, 1971); and arguments of volition and commitment predicates (*wish; offer*) are presented as possible in a future point in time.

Modal Particles. These include modal auxiliaries (*could, may, must*), but also clausal and sentential adverbial modifiers (*maybe, likely, possibly*).

Polarity Particles. These include elements of a varied nature: adverbs (*not, until*), quantifiers (*no, none*), pronouns (*nobody*), etc. They switch the polarity of its context. When scoping over a modal particle, they also affect its modal interpretation.

2.4.2. Syntactic Contexts

Syntactic structures conveying factuality information involve two clauses, one embedded under the other. In some cases, the embedded event is presupposed as holding; e.g., relative clauses (19), cleft sentences (20), and subordinated temporal clauses.

(19) *Rice, who became secretary of state two months ago today, took stock of a period of tumultuous change.*

(20) *It was Mr. Bryant who, on July 19, 2001, asked Rep. Bartlett to pen and deliver a letter to him.*

In others, the event denoted by the embedded clause is intensional in nature; e.g., purpose clauses (21) and conditional constructions (22).

(21) *The environmental commission must adopt regulations to ensure people are not exposed to radioactive waste.*

(22) *EZLN will return to the negotiating table if the conflict zone is* **demilitarized**.

3. Challenges in identifying event factuality

Annotating event factuality poses challenges at two levels. First, factuality is in many cases the result of different factuality markers interacting among them. They can all be in the local context of the event, but it is also common for them to be at different levels. Second, the factuality of an event is always relative to one or more sources. Hence, they must be included as part of the annotation scheme as well. The following subsections elaborate on these two issues. Refer to Saurí (2008) for a more comprehensive view on event factuality and its identification.

3.1. Interpreting the factuality of events

Event factuality involves local but also non-local information. Consider the following examples:[4]

(23) 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.

The event *inspections* in (23a), where *allow* is embedded under the factive predicate *continue*, is characterized as a fact in the world –i.e., there have been such inspections. Example (23b), 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 (23c), 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 obtained from the strict local modality and polarity operators scoping over that event but, if present, appealing to their interaction with other non-local markers as well. Consequently, annotating factuality from a surface-based approach, accounting for the structural elements and without considering their interaction, will miss an important piece of information.

3.2. Relevant sources

The second 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 question answering or narrative understanding. For example, event e in (24) (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 the assertion of a specific source; namely, Milosevic's son.

(24) 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 predicates of reporting (*say, tell*), knowledge and opinion (e.g., *believe, know*), psychological reaction (*regret*), etc. Because of their role in introducing a new source, we call them Source Introducing Predicates (SIPs).

The status of the additional sources is, however, different from that of the author of the text. For instance, in (25) 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, for that matter, to the assessment of G-7 leaders about e_3.

(25) Izvestiya **said**$_{e_1}$ that the G-7 leaders **pretended**$_{e_2}$ everything **was OK**$_{e_3}$ in Russia's economy.

Thus, we need to appeal to the notion of *nested source* as presented in Wiebe et al. (2005). *Izvestiya* is not a licit source of the factuality of event e_2, but *Izvestiya* according to the author instead, represented here as *izvestiya_author*.[5]

[4] As startling as it may result, the original sentence in this set is (23b), from the BNC.

[5] Equivalent to the notation $<author,izvestiya>$ in Wiebe's work.

Similarly, the source referred to by the G-7 leaders corresponds to the chain: *g7leaders_izvestiya_author*.
As it happens, the same event can have more than one relevant source relative to which its factuality is assessed. In some cases, they coincide in the factual status of the event but in others there is disagreement. In (25), for example, event e_3 is assessed as being a fact (CT+) according to the G-7 leaders (corresponding to source *g7leaders_izvestiya_author*), but as being false (CT−) according to Izvestiya (i.e., *izvestiya_author*). The text author, on the other hand, remains uncommitted (Uu).
The factuality value assigned to events in text must be relative to the relevant sources at play, which may be one or more. Only under this assumption it is possible to account for the potential contradictions between factual values assigned to the same event, and the different opinions commonly found in news reports.

4. Factuality information in existing corpora

To our knowledge, factuality-related information is annotated in three corpora: the MPQA Opinion Corpus (Wiebe et al., 2005), the Penn Discourse TreeBank (Miltsakaki et al., 2004), and TimeBank (Pustejovsky et al., 2003). Currently, it is also being annotated in the ACE 2008 program.[6]
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 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 perspective (i.e., objective vs. subjective).
Another common feature between the MPQA Opinion Corpus scheme and our model of event factuality is the encoding of sources. Both approaches structure them as chains of nested sources. From our perspective, however, the MPQA Opinion Corpus is limited in that it only acknowledges one relevant source for each event.
Another limitation in the MPQA annotation scheme is that it is not grammatically grounded. That is, the annotation of text spans is not guided according to the grammatical structure of the sentence, and this can pose an obstacle for tasks of automatic recognition.
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 to them (Prasad et al., 2007). The task is approached from a compositional approach, contrary to the MPQA Opinion Corpus.
In spite of these similarities, there are two significant differences. With regard to sources, PDTB does not encode the nesting relation that can hold among them, neither accounts for the possibility of more than one source for a given abstract object (or event).
The second difference concerns the factuality degree associated to the attributed event, which is assigned based on the type of action described by the predicate embedding it. In particular, events embedded under communication predicates are characterized as asserted; events embedded by propositional attitude predicates, as beliefs; and events embedded under factive predicates, 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 a nuance of belief. Similarly, *forget* is a factive predicate which, contrary to others in its class, expresses an uncommitted (or ignorant) stance of the source (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. Finally, PDTB annotation is not concerned with the effect of other markers of modality (modal auxiliaries and adverbials) on the factuality of abstract objects.
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. Given the surface-based approach of TimeML, TimeBank is the corpus that takes the most compositional approach to annotation among the three reviewed corpora.
The factuality-relevant information encoded in TimeBank is mainly lexical: grammatical particles expressing event modality and polarity, as well as event selecting predicates (*cf.* section 2.4.1.), which project a factual value to their embedded event by means of subordination links (or slinks). Thus, TimeBank provides us with the basic components expressing factuality information in text –a consequence of the explicit surface-based approach of TimeML. And 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. Creating a corpus of event factuality

5.1. FactBank

FactBank is a corpus annotated with factuality information. It consists of 208 documents and contains a total of 8837 events manually annotated. FactBank includes all the documents in TimeBank and a subset of those in the AQUAINT TimeML Corpus (A-TimeML Corpus)[7]. The contribution of each of these corpora to FactBank is shown in Table 3.

Table 3: FactBank sources

	# Documents		# Events	
TimeBank	183	(88%)	7935	(90%)
A-TimeML Corpus	25	(12%)	902	(10%)
Total	**208**		**8837**	

Because both TimeBank and AQUAINT TimeML Corpus are annotated with the TimeML spec, FactBank incorporates a second layer of factuality information on top of that in the original corpora. Thus, while the former two encode the structural elements expressing factuality information in language, the latter represents the resulting interpretation. The new annotation is kept in separate documents

[6] http://projects.ldc.upenn.edu/ace/annotation/. Because it still is an ongoing project, we will not comment on that corpus here.

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

and is linked to the original data by means of the events IDs, which are the same in both annotation layers.[8]

5.2. Corpus annotation

We argued earlier that identifying event factuality requires linguistic processing at different layers. First, it involves the interaction of local and non-local context. Second, it puts into play at least one, but generally more, relevant sources for each event, which bear a nesting relation among them. Hence, if not structured adequately, the annotation task could become too complex and would inevitable result in a questionable outcome. Annotating event factuality needs to be addressed by steps that could 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.

5.2.1. Task 1: Identifying Source-Introducing Predicates (SIPs)

Given a text with the events already recognized and marked as such, the annotators identified those that correspond to Source-Introducing Predicates. SIPs were briefly described in section 3.2. 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, will have a role in assessing the factuality of the SIP event complement –recall example (25).

This initial task allowed annotators to get familiarized with both the notion of source and the notion of SIP as marker of factuality information. Moreover, for processing purposes Saurí & Pustejovsky (2007) show that identifying SIPs is fundamental for the automatic computation of relevant sources. The manual annotation resulting from this task was then used to prepare the final task.

5.2.2. Task 2: Identifying sources

The annotator was provided with a text with the following information already annotated: (a) all the SIPs in the text –obtained from the previous task; 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 syntactic functions listed here:[9]

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***.)[10]
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*).

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 posterior 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 then interpreted as generic –i.e., it can be paraphrased as *everybody*. They will be represented as GEN in the resulting chain expressing the relevant source (e.g., GEN_*author*).

5.2.3. Task 3: Assigning factuality values

This final task was devoted to selecting the factuality value assigned to events by each of their relevant sources. 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 them presented in Saurí & Pustejovsky (2007).

The annotators had to choose among the set of factuality values presented in Table 4, which corresponds *grosso modo* to Table 1 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 they were also considered unnecessary by the annotators when looking at 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 section 2.3.

6. Evaluation

FactBank has been annotated by a pair of annotators. Overall, three annotators participated in the effort: annotators A and B participated in the first task, and annotators B and C carried out tasks 2 and 3. All of them are competent undergraduate Linguistics Majors. In addition, there were two

[8] FactBank annotation can be expressed by means of XML tags representing the factuality value assigned by a source to a given event. Because each event can be assigned more than one factuality value (as many as relevant sources it has), these must be non-consuming tags. Alternatively, given the correspondence between events IDs in both layers, the mapping can be established by means of stand-off markup as well.

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

[10] In this and coming examples, the new source candidate is marked in bold face and the SIP, underlined.

Table 4: Factuality values

VAL	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.

adjudicators handling cases of disagreement in each task before annotators would continue with the next one.

Task 1. The interannotation ratio achieved is k=0.88 over 40% of the corpus (on the number of events).[11] Some of the most common cases of disagreement concern:

- SIP candidates with implicit sources –e.g., generic, as in: *He's **expected** to meet with Iraqi deputy prime minister Tariq Aziz later this afternoon.*
- SIP candidates lacking an explicit event complement (e.g., *The executives didn't **disclose** the size of the expected gain.*).
- Negated SIP candidates (e.g., *didn't* **disclose***, did not* **tell**, in the examples above).

Task 2. The interannotation agreement achieved for this task is k=0.95 over 40% of the corpus (on the number of events). Such good results come as no surprise since it is a very well-defined task, both in syntactic and semantic terms –essentially, it requires identifying SIP logical subjects. The most common cases of disagreements are those in which:

- There is a second expression in the text correfering with the new source. For example, the first person pronoun in a quoted fragment (e.g., *"We are going to maintain our forces in the region for the foreseeable future,"* **said** *spokesman Kenneth* **Bacon**.)[12]

  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 referred to a non-human entity (e.g., **Reports** *attributed to the Japanese foreign ministry **said** ...*). One of the annotators would choose a different option.

Task 3. Interannotion agreement for this last task scores at k=0.82 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. In a comparable work devoted to classify certainty in text according to a five-fold categorization (*absolute, high, moderate, low*, and *uncertain*) (Rubin, 2007), the interannotation score obtained was k=0.15, which improved to k=0.41 when stricter annotation instructions were provided.

Furthermore, an analysis of disagreement cases on the 10% of our corpus shows that around two thirds of them are cases of true ambiguity, originated from different constructions. Some of the most common concerned the scope of a reporting predicate –or, in other words, the span of the attributed fragment. In (26), for example, the reporting predicate (in bold face) can be interpreted as scoping over both events *want* and *traveled*, or only only over *traveled*.

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

A second common case of ambiguity is caused by syntactic constructions typically triggering a presupposition (e.g., relative clauses, temporal clauses, appositions) when embedded under a reporting predicate (27). Annotators would disagree on whether the presupposition would be projected to the main clause –in our terms, the disagreement concerns whether the author of the text commits to the embedded event (underlined below) as a fact.

(27) 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 and some campaign advisers.

7. Conclusions

Event factuality is an important component for representing events in discourse, but identifying it poses a two-fold challenge. First, factuality is in many cases the result of different factuality markers interacting among them. They can all be in the local context of the event, but it is also common for them to be at different levels. Second, the factuality value assigned to events in text must be relative to the relevant sources at play, which may be one or more.

In this paper, we introduced FactBank, a corpus of events annotated with factuality. FactBank contributes a semantic layer of factuality information on top of the grammar-based layer provided in TimeBank.

The interannotation agreement scores obtained for the three annotation tasks we designed are encouraging. Specifically, for the task of selecting the factuality value assigned to events by each of their relevant sources, we achieved k=0.82 over 30% of the corpus. That suggests that event factuality as modeled in our work is well-grounded in linguistic data, and that its identification is achievable using an approach along the lines of that proposed here. FactBank will be made available to the community in a near future.

[11] We apply Cohen *Kappa* (Cohen, 1960), hence assuming any potential distortion in the resulting figures due to the skewed distribution of categories (the so-called prevalence problem) as well as the degree to which the annotators disagree (the bias problem). Refer to Di Eugenio & Glass (2004).

[12] In this and the following examples, the SIP is presented in bold face and the new source to be selected in bold face and underlined. If an additional expression enters in consideration as new source candidate as well, it will only be underlined.

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*: 86–90.

Bergler, S. (1992). *Evidential Analysis or Reported Speech*. PhD thesis, Brandeis University.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). Building a discourse-tagged corpus in the framework of rhetorical structure theory.

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

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

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). 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*.

Halliday, M. A. K. & Matthiessen, C. M. (2004). *An introduction to functional grammar*. London: 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*: 171–176.

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

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

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

Karttunen, L. (1971). Some observations on factivity. *Papers in Linguistics*, *4*, 55–69.

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

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

Lewis, D. (1968). Counterpart theory and quantified modal logic. *Journal of Philosophy*, *65*, 113–126.

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*: 17–24.

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

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

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).

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*.

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).

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., & Lazo, M. (2003). The TimeBank corpus. In *Proceedings of Corpus Linguistics 2003*, (pp. 647–656).

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.

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*.

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 1st IEEE International Conference on Semantic Computing*.

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

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

Verhagen, M., Stubbs, A., & Pustejovsky, J. (2007). Combining independent syntactic and semantic annotation schemes. In *Proceedings of the Linguistic Annotation Workshop*.

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

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