

FactBank 1.0

Corpus documentation

Roser Saurí James Pustejovsky

June 11, 2009

1. **Corpus name:** FactBank 1.0.

2. **Authors**

Roser Saurí (contact person)

email: roser@cs.brandeis.edu, roser.sauri@gmail.com

phone: +1-781-736-2745

James Pustejovsky

email: jamesp@cs.brandeis.edu

phone: +1-781-736-2709

3. **Data type:** Text.

4. **Languages:** English (eng).

5. **Description of the corpus and suggested uses**

FactBank is a corpus of news reports in which event mentions are annotated with their degree of factuality, expressing whether they correspond to actual situations in the world, situations that have not happened, or situations of uncertain interpretation.

Event factuality is not an inherent feature of events but a matter of perspective. Different discourse participants may present divergent views about the factuality of the very same event. Consequently, in FactBank the factuality degree of events is assigned relative to the relevant sources at play. In this way, it can adequately reflect the divergence of opinions regarding the factual status of events, as is common in news reports.

The annotation language is grounded on established linguistic analyses of the phenomenon, which facilitated the creation of a battery of discriminatory tests for distinguishing between factuality values. Furthermore, the annotation procedure was carefully designed and divided into basic, sequential annotation tasks. This made it possible for hard tasks to be built on top of simpler ones, while at the same time allowing annotators to become incrementally familiar with the complexity of the problem. As a result, FactBank annotation achieved a relatively high interannotation agreement, kappa=0.81, a positive result when considered against similar annotation efforts.

FactBank has been built on top of TimeBank 1.2 and a fragment of AQUAINT TimeML, two corpora annotated with the TimeML specification language (Pustejovsky et al., 2005).

This resulted in a double-layered annotation of event factuality: where TimeBank 1.2 and AQUAINT TimeML encode most of the basic structural elements expressing factuality information, FactBank represents the resulting factuality interpretation. 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, a component fundamental in tasks requiring some degree of text understanding, such as Textual Entailment, Question Answering, or Narrative Understanding.

For an exhaustive description of the corpus, its design, building process, and results obtained, see Saurí & Pustejovsky (2009). Furthermore, the annotation guidelines (Saurí, 2008a) can be found in the documentation directory of the current release.

6. Data sources

FactBank includes all the documents in TimeBank 1.2 (released through LDC, see Pustejovsky et al. (2006)) and a subset of those in the AQUAINT TimeML Corpus (refer to <http://www.timeml.org/site/timebank/timebank.html>).

- Data from TimeBank 1.2:

Some of the documents are from the DUC1 summarization evaluation that NIST ran in 2001 (files whose names start with “AP”, “LA” “SJM”, and also of the “WSJ” files). The rest of the articles are from ACE corpora. The file names that start with “ABC”, “CNN”, “PRI”, “VOA”, “ea”, and “ed” are broadcast news. Those that start with “APW” and “NYT” are newsire. All of these are included in LDC catalog item LDC2003T11. The other ACE corpus is WSJ. The documents are included in LDC catalog item LDC99T42.

- Data from AQUAINT TimeML:

The documents used in FactBank were taken from two topics from the TREC novelty track (see <http://trec.nist.gov/tracks.html>):

- N35 NATO, Poland, Czech Republic, Hungary
- N45 Slepian abortion murder

7. Rough size of the corpus

FactBank consists of 208 documents (183 articles from TimeBank and 25 articles from AQUAINT TimeML) and has over 77,000 tokens (including punctuation marks). Furthermore, it contains around 9,500 events that were identified by manual annotation.

8. Project for which the corpus was developed

FactBank was developed as part of the doctoral research presented in Saurí (2008b). Such research was funded by two grants to Prof. James Pustejovsky for the projects:

- Temporal Awareness and Reasoning Systems for Question Interpretation (TARSQI): ARDA Contract No. NBCHC040027.
- Tarsqi Toolkit (TTK): NAVAIR Contract No. N61339-06-C-0140.

9. Corpus structure and data attributes

All FactBank markup is standoff, represented through a set of 20 tables which can be easily loaded into a DB. Each table is contained in an independent text file, where fields are separated by three consecutive bars (i.e., |||). The data in fields of string type are presented between simple quotations (').

The following lists each table file and its size in bytes, by alphabetical order.

8569	fb_corefSource.txt
507701	fb_event.txt
1223419	fb_factValue.txt
205594	fb_relSource.txt
125467	fb_sipAndSource.txt
93364	fb_sip.txt
212490	fb_source.txt
71284	fb_sourceString.txt
7419	files.txt
7538984	offsets.txt
520197	sentences.txt
32431	tml_alink.txt
539557	tml_event.txt
856453	tml_instance.txt
41929	tml_signal.txt
361375	tml_slink.txt
223987	tml_timex3.txt
957233	tml_tlink.txt
3460463	tokens_ling.txt
4190086	tokens_tml.txt

FactBank has been built on top of TimeBank 1.2 and AQUAINT TimeML, both of which are marked up with inline XML-based annotation. Hence, in order to facilitate data compatibility, the current FactBank release contains the TimeBank 1.2 and AQUAINT TimeML annotation in standoff, table-based format as well.

Files which name begins by the prefix 'tml_' contain annotation specific of TimeBank 1.2 and AQUAINT TimeML. Files began by the prefix 'fb_' contain the FactBank annotation. The remaining 5 files (*files.txt*, *offsets.txt*, *sentences.txt*, *tokens_ling.txt*, and *tokens_tml.txt*) present general information regarding the documents, sentences, tokens, and offsets in the corpus.

In what follows, we describe the structure of each table (attribute names and suggested primary keys). Furthermore, each attribute will be accompanied by a brief description of its contents and its data type (e.g., string, numeric). Attributes with the same name across tables refer to the same information. We will explicitly indicate what attributes from different tables refer to the same information but nevertheless do not share the same name.

A. Tables with general information:

Table: *files*. Containing the file name of each document in the corpus, together with its provenance, namely, TimeBank 1.2 or AQUAINT TimeML.

- *Data attributes:*
 - `file` String. Name of the corpus document.
 - `corpus` String. Provenance of the document.
- *Primary key:* `file`.

Table: *sentences*. Containing data relative to each sentence in the corpus documents.

- *Data attributes:*
 - `file` String. Name of the corpus document.
 - `sentId` Numeric. ID of the sentence where the string belongs to.
 - `sent` String. Sentence text.
- *Primary keys:* `file`, `sentId`.

Table: *tokens_ling*. Linguistic information relative to each sentence token. Tokens include words and punctuation marks, disregarding blank spaces and text relative to the document metadata.

- *Data attributes:*
 - `file` String. Name of corpus document.
 - `sentId` Numeric. ID of the sentence where the token belongs to.
 - `tokLoc` Numeric. Location of the token in the sentence.
 - `text` String. Token text.
 - `pos` String. Part of speech of the token.
- *Primary keys:* `file`, `sentId`, `tokLoc`.

Table: *tokens_tml*. Information relative to the TimeML annotation (either from TimeBank 1.2 or AQUAINT TimeML) of each token in a document sentence. As in table *tokens_ling*, tokens include words and punctuation marks, disregarding blank spaces and text relative to the document metadata.

- *Data attributes:*
 - `file` String. Name of corpus document.
 - `sentId` Numeric. ID of the sentence where the token belongs to.
 - `tokLoc` Numeric. Location of the token in the sentence.
 - `text` String. Token text.
 - `tmlTag` String. TimeML entity tag (i.e., EVENT, TMEX3, SIGNAL) associated to the token.
 - `tmlTagId` String. Tag ID as assigned during the TimeML annotated

corpora, that is, TimeBank 1.2 and AQUAINT TimeML.

- `tmlTagLoc` String. Location of the token in the tag (B: beginning, I: inside).
This information is necessary for tags containing multiple tokens.

Tokens not annotated as TimeML entities will have the last 3 attributes empty.

- *Primary keys:* `file`, `sentId`, `tokLoc`.

Table: *offsets*. Containing the offsetsof token and non-token strings in each corpus document.

- *Data attributes:*

- `file` String. Name of the corpus document.
- `offsetInit` Numeric. Initial offset of the string.
- `offsetEnd` Numeric. Offset at the end of the string.
- `sentId` Numeric. ID of the sentence where the string belongs to.
- `tokLoc` Numeric. Location of the string in the sentence.
- `text` String. Text of the string.

Due to historical reasons, attributes `sentId` and `tokLoc` are assigned only to text tokens (words and punctuation marks), and disregard blank spaces and spurious text (e.g., relative to the document metadata). Hence, in these two attributes, non-token strings receive the value NULL.

- *Primary keys:* `file`, `offsetInit`.

B. Tables with the TimeML annotation of the documents. They contain the annotation in all TimeBank 1.2 and some documents from AQUAINT TimeML, in stand-off format.

Table: *tml_timex3*. Containing the annotation of TIMEX3 entities in TimeBank 1.2 and AQUAINT TimeML. Refer to the TimeML documentation (Pustejovsky et al., 2005).

- *Data attributes:*

- `file` String. Name of the corpus document.
- `sentId` Numeric. Sentence ID.
- `tId` String. Timex3 ID.
- `type` String. Timex3 type attribute.
- `value` String. Timex3 value attribute.
- `mod` String. Timex3 mod attribute.
- `functionInDoc` String. Timex3 functionInDoc attribute.
- `temporalFunction` Boolean (true, false). Timex3 temporalFunction attribute.
- `anchorTimeId` String. Timex3 anchorTimeId attribute.
- `beginPoint` String. Timex3 beginPoint attribute.
- `endPoint` String. Timex3 endPoint attribute.

- `freq` String. Timex3 `freq` attribute.
- `quant` String. Timex3 `quant` attribute.
- `timexText` String. Text included in the span of the Timex3 tag.
- `tokenText` String. Tokens which are at least partially included in the span of the Timex3 tag. This attribute is added for cases where the Timex3 tag expands only over a part of a token. For example:

<u><code>tokenText:</code></u>	<u><code>timexText:</code></u>
- <i>PRI19980121</i>	- <i>19980121</i>
- <i>eight-year-long</i>	- <i>eight year</i>

In most cases, however, the information in this attribute is the same as in the previous one, `timexText`.

- *Primary keys:* `file`, `tId`.

Table: *tml_event*. Containing the annotation of EVENT entities in TimeBank 1.2 and AQUAINT TimeML. Refer to Pustejovsky et al. (2005).

- *Data attributes:*

- `file` String. Name of the corpus document.
- `sentId` Numeric. Sentence ID.
- `eId` String. Event ID.
- `eClass` String. Event `class` attribute.
- `eText` String. Text included in the Event tag.

- *Primary keys:* `file`, `eId`.

Table: *tml_instance*. Containing the annotation of INSTANCE entities in TimeBank 1.2 and AQUAINT TimeML. See Pustejovsky et al. (2005).

- *Data attributes:*

- `file` String. Name of the corpus document.
- `eId` String. Event ID.
- `eiId` String. Event instance ID.
- `tense` String. Event tense.
- `aspect` String. Event aspect.
- `pos` String. Part of speech of the event expression.
- `polarity` String. Event polarity.
- `modality` String. Event modality.
- `cardinality` String. Event cardinality.
- `signalId` String. ID of the signal related to the event instance, if applicable.

- *Primary keys:* `file`, `eiId`.

Table: *tml_signal*. Containing the annotation of SIGNAL entities in TimeBank 1.2 and AQUAINT TimeML.

- *Data attributes:*

- **file** String. Name of the corpus document.
- **sentId** Numeric. Sentence ID.
- **signalId** String. ID of the Signal tag.
- **signalText** String. Text included in the span of the Signal tag.
- **tokenText** String. Tokens which are at least partially included in the span of the Signal tag. This attribute is added for cases where the Signal tag expands only over a part of a token. For example:

<u>tokenText:</u>	<u>signalText:</u>
- overnight	- over

In most cases, however, the information in both attributes is the same.

- *Primary keys:* **file**, **signalId**.

Table: *tml_tlink*. Containing the annotation of TLINK entities in TimeBank 1.2 and AQUAINT TimeML.

- *Data attributes:*

- **file** String. Name of the corpus document.
- **lId** String. TLink ID.
- **eId_1** String. ID of event 1 in the TLink. Same information as attributes named **eId** in other tables.
- **eId_2** String. ID of event 2 in the TLink. Same information as attributes named **eId** in other tables.
- **eiId_1** String. Event instance ID of event 1 in the TLink. Same information as attributes named **eiId** in other tables.
- **eiId_2** String. Event instance ID of event 2 in the TLink. Same information as attributes named **eiId** in other tables.
- **tId_1** String. ID of Timex3 1 in the TLink. Same information as attributes named **eId** in other tables.
- **tId_2** String. ID of Timex3 2 in the TLink. Same information as attributes named **eId** in other tables.
- **relType** String. Relation type of the TLink (e.g., 'before', 'after', 'during', 'includes', 'is_included', etc.).
- **signalId** String. ID of the element signaling the relation encoded by the TLink, if any.
- **eText_1** String. Text of event 1 in the TLink. Same information as attributes named 'eText' in other tables.
- **eText_2** String. Text of event 2 in the TLink. Same information as attributes

named 'eText' in other tables.

- `signalText` String. Text of the element signaling the relation encoded by the TLink, if any.

- *Primary keys:* `file`, `eiId_1`, `eiId_2`, `tId_1`, `tId_2`.

Table: *tml_alink*. Containing the annotation of ALINK entities in TimeBank 1.2 and AQUAINT TimeML.

- *Data attributes:*

- `file` String. Name of the corpus document.
- `lId` String. ALink ID.
- `eId_1` String. ID of event 1 in the ALink. Same information as attributes named `eId` in other tables.
- `eId_2` String. ID of event 2 in the ALink. Same information as attributes named `eId` in other tables.
- `eiId_1` String. Event instance ID of event 1 in the ALink. Same information as attributes named `eiId` in other tables.
- `eiId_2` String. Event instance ID of event 2 in the ALink. Same information as attributes named `eiId` in other tables.
- `relType` String. Relation type of the ALink (e.g., 'initiates', 'continues', 'culminates', etc.)
- `signalId` String. ID of the element signaling the relation encoded by the ALink, if any.
- `eText_1` String. Text of event 1 in the ALink. Same information as attributes named `eText` in other tables.
- `eText_2` String. Text of event 2 in the ALink. Same information as attributes named `eText` in other tables.
- `signalText` String. Text of the element signaling the relation encoded by the ALink, if any.

- *Primary keys:* `file`, `eiId_1`, `eiId_2`.

Table: *tml_slink*. Containing the annotation of SLINK entities in TimeBank 1.2 and AQUAINT TimeML.

- *Data attributes:*

- `file` String. Name of the corpus document.
- `lId` String. SLink ID.
- `eId_1` String. ID of event 1 in the SLink. Same information as attributes named `eId` in other tables.
- `eId_2` String. ID of event 2 in the SLink. Same information as attributes named `eId` in other tables.
- `eiId_1` String. Event instance ID of event 1 in the SLink. Same information as attributes named `eiId` in other tables.

- `eiId_2` String. Event instance ID of event 2 in the SLink. Same information as attributes named `eiId` in other tables.
 - `relType` String. Relation type of the SLink (e.g., ‘evidential’, ‘modal’, ‘factive’, ‘counter_factive’, etc.)
 - `signalId` String. ID of the element signaling the relation encoded by the SLink, if any.
 - `eText_1` String. Text of event 1 in the SLink. Same information as attributes named `eText` in other tables.
 - `eText_2` String. Text of event 2 in the SLink. Same information as attributes named `eText` in other tables.
 - `signalText` String. Text of the element signaling the relation encoded by the SLink, if any.
- *Primary keys:* `file`, `eiId_1`, `eiId_2`.

C. Tables containing the annotation specific from FactBank:

Table: *fb_event*. Events in FactBank that had their factuality degree annotated. They correspond to the events in the TimeML-based annotation presented in table *tml_event*. However, the two tables differ in that in *fb_event*, different instances of the same event are listed independently and distinguished by means of the attribute `eiId`, whereas in the TimeML-based annotation, event instances are stored in an table independent from that of events, namely *tml_instance*.

The annotation of events in FactBank goes along the lines of the philosophy adopted in the TimeML ISO specification, which does not support the `INSTANCE` tag.

- *Data attributes in the current table:*
 - `file` String. Name of the corpus document.
 - `sentId` Numeric. Sentence ID where the event string appears.
 - `eId` String. Event ID.
 - `eiId` String. Event instance ID.
 - `eText` String. Event text.
- *Primary keys:* `file`, `eId`, `eiId`.

Table: *fb_sourceString*. Containing those sources that are introduced in discourse by Source-Introducing Predicates (SIPs), including as what we refer to as ‘generic sources’ as well. Generic sources are present in contexts such as impersonal constructions (e.g., ‘*it was expected that...*’), and thus do not consume any textual string.¹

- *Data attributes:*

¹In that sense, the name of the table can be a bit confusing.

- `file` String. Name of the corpus document.
- `sentId` Numeric. Sentence ID.
- `sourceId` String. Source ID.
- `sourceLoc` Numeric. Position in text of the token expressing the string. Same information as attribute `tokLoc` in tables *offsets*, *token_ling*, and *token_tml*.
- `sourceText` String. Textual string denoting the source.

In the table, the values for the attributes `sourceLoc` and `sourceText` for generic sources have been arbitrarily stipulated, since they do not consume any textual string and hence do not occupy a particular position (`sourceLoc`) in the sentence.

generic: `sourceLoc=-2, sourceText=GEN`

- *Primary keys:* `file`, `sentId`, `sourceId`.

Table: *fb_source*. Containing the whole set of sources that play a role in evaluating the events expressed in each corpus sentence. Hence, it includes (a) sources introduced in discourse by SIP (and therefore already present in table *fb_sourceStrings*), as well as (b) the text author and (c) what we call 'dummy sources'. Dummy sources are relevant for sentences that are clearly reported by somebody (because they are presented between quotation marks), but make no explicit mention of who that source is. For example:

“The main thing is to stop the oil from coming out of there.”

Dummy sources would not be needed if source identification was not constricted at the sentence level.

- *Data attributes:*

- `file` String. Name of the corpus document.
- `sentId` Numeric. Sentence ID.
- `sourceId` String. Source ID.
- `sourceLoc` Numeric. Position in text of the token expressing the string. Same information as attribute `tokLoc` in tables *offsets*, *token_ling*, and *token_tml*.
- `sourceText` String. Textual string denoting the source.

In the table, the text author source received the default ID of `s0`. In addition, the values for the attributes `sourceLoc` and `sourceText` for sources that consume no textual string in the text (i.e., author, generic, and dummy sources) are arbitrarily assigned as follows:

author: `sourceLoc=-1, sourceText=AUTHOR`

generic: `sourceLoc=-2, sourceText=GEN`

dummy: `sourceLoc=-3, sourceText=DUMMY`

- *Primary keys:* `file`, `sentId`, `sourceId`.

Table: *fb_corefSource*. Containing the (automatically identified) coreferences between two different strings that denote a factuality source. Correfering sources are identified by unique IDs (attribute `corefSourceId`), which are formed by conjoining the IDs of each independent source string (e.g., `s4` and `s0`) separated by = (e.g., `s4=s0`).

- *Data attributes in the current table:*
 - `file` String. Name of the corpus document.
 - `sentId` Numeric. Sentence ID where the coreferring strings were found.
 - `corefSourceId` String. ID of the source coreference.
 - `sourceId_1` String. ID of source string 1. Same information as attributes named `sourceId` in other tables.
 - `sourceId_2` String. ID of source string 2. Same information as attributes named `sourceId` in other tables.
 - `sourceLoc_1` Numeric. Position in text of the source string 1. Same information as attribute `sourceLoc` in other tables, as well as attribute `tokLoc` in tables *offsets*, *token_ling*, and *token_tml*.
 - `sourceLoc_2` Numeric. Position in text of the source string 2. Same information as attribute `sourceLoc` in other tables, as well as attribute `tokLoc` in tables *offsets*, *token_ling*, and *token_tml*.
 - `sourceText_1` String. Textual string denoting source 1. Same information as attributes named `sourceText` in other tables.
 - `sourceText_2` String. Textual string denoting source 2. Same information as attributes named `sourceText` in other tables.
- *Primary keys:* `file`, `corefSourceId`.

Table: *fb_relSource*. Containing the sources that are relevant for evaluating the factuality degree of each event in text.

- *Data attributes:*
 - `file` String. Name of the corpus document.
 - `sentId` Numeric. Sentence ID where the relevant source appears.
 - `relSourceId` String. Relevant source ID.
 - `relSourceText` String. Relevant source text.
- *Primary keys:* `file`, `sentId`, `relSourceId`.

Table: *fb_factValue*. Containing the factuality degree values assigned to each event in text according to a relevant information source. Factuality values belong to a closed set of 8 (e.g., certain (abbreviated as: `CT+`), probable (`PR+`), possible (`PS+`), etc.), as presented in detail in Saurí and Pustejovsky (2009).

- *Data attributes:*

- `file` String. Name of the corpus document.
- `sentId` Numeric. Sentence ID where the assessed event appears.
- `fvId` String. Factuality judgment ID.
- `eId` String. Event ID.
- `eiId` String. Event instance ID.
- `relSourceId` String. ID of the relevant source according to which the factuality degree of the event is assessed.
- `eText` String. Event text.
- `relSourceText` String. Relevant source text.
- `factValue` String. Factuality value assigned to the event by the relevant source.

- *Primary keys:* `file`, `fvId`.

Table: *fb_sip*. Containing information relative to the Source-Introducing Predicates (SIPs). This table is information not strictly necessary regarding the factuality degree of events. It contains the information resulting from Annotation Task 1 (refer to Saurí & Pustejovsky (2009)), and hence we estimated it could be of interest to the public as well.

- *Data attributes:*

- `file` String. Name of the corpus document.
- `sentId` Numeric. Sentence ID.
- `sip_eId` String. Event ID of the SIP. Same information as attributes named `eId` in other tables.
- `sip_eiId` String. Event instance ID of the SIP. Same information as attributes named `eiId` in other tables.
- `sip_Text` String. Event text. Same information as attributes named `eText` in other tables.

- *Primary keys:* `file`, `sip_eiId`.

Table: *fb_sipAndSource*. Containing the pairs of Source-Introducing Predicates (SIPs) and the source they introduce in discourse. Refer to Saurí & Pustejovsky (2009). This table is information not strictly necessary regarding the factuality degree of events. It contains the information resulting from Annotation Task 2. We estimate it could be of interest to the public as well.

- *Data attributes:*

- `file` String. Name of the corpus document.
- `sentId` Numeric. Sentence ID.
- `sip_eId` String. Event ID of the SIP. Same information as attributes named `eId` in other tables.
- `sip_eiId` String. Event instance ID of the SIP. Same information as attributes named `eiId` in other tables.
- `sip_Text` String. Event text. Same information as attributes

named `eText` in other tables.

- `sourceId` String. ID of source introduced by the SIP.
- `sourceText` String. Textual string denoting the source.

- *Primary keys:* `file`, `sip_eiId`.

10. Directory structure

`doc/` Documentation related to the present release. Containing: this *readme.pdf* file as well as the *annotation guidelines* used for annotating the corpus.

`data/`

`annotation/` Files containing the corpus annotation. Each file corresponds to one of the tables presented above.

`original/` Files containing the original documents constituting the corpus.

11. Copyright

The annotations in this data collection are copyrighted by Brandeis University. User acknowledges and agrees that: (i) as between User and Brandeis University, Brandeis University owns all the right, title and interest in the Annotated Content, unless expressly stated otherwise; (ii) nothing in this Agreement shall confer in User any right of ownership in the Annotated Content; and (iii) User is granted a non-exclusive, royalty free, worldwide license (with no right to sublicense) to use the Annotated Content solely for academic and research purposes. This Agreement is governed by the law of the Commonwealth of Massachusetts and User agrees to submit to the exclusive jurisdiction of the Massachusetts courts.

Note: The textual news documents annotated in this corpus have been collected from a wide range of sources and are not copyrighted by Brandeis University. The user acknowledges that the use of these news documents is restricted to research and/or academic purposes only.

References

- 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., 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>).
- Saurí, R. (2008a). *FactBank 1.0. Annotation Guidelines*. Brandeis University. Available at: http://www.cs.brandeis.edu/~roser/pubs/fb_annotGuidelines.pdf.
- Saurí, R. (2008b). *A Factuality Profiler for Eventualities in Text*. PhD thesis, Brandeis University. Available at: http://www.cs.brandeis.edu/~roser/pubs/sauriDiss_1.5.pdf.
- Saurí, R. & Pustejovsky, J. (2009). FactBank. A corpus annotated with event factuality. *Language Resources and Evaluation*. (Online version: <http://www.springerlink.com/content/mk122157x3113678/?p=20ede75cdad340dfb76d33a26089ef63&pi=0>).