The EVALITA 2016 Event Factuality Annotation Task (FactA)

Anne-Lyse Minard[1,3], Manuela Speranza[1], Tommaso Caselli[2]
[1] Fondazione Bruno Kessler, Trento, Italy
[2] VU Amsterdam, the Netherlands
[3] Dept. of Information Engineering, University of Brescia, Italy
{speranza,minard}@fbk.eu
t.caselli@vu.nl

Abstract

English. This report describes the FactA (Event Factuality Annotation) Task presented at the EVALITA 2016 evaluation campaign. The task aimed at evaluating systems for the identification of the factuality profiling of events. Motivations, datasets, evaluation metrics, and post-evaluation results are presented and discussed.

Italiano. *Questo report descrive il task di valutazione FactA (Event Factaulity Annotation) presentato nell'ambito della campagna di valutazione EVALITA 2016. Il task si prefigge lo scopo di valutare sistemi automatici per il riconoscimento della fattualitá associata agli eventi in un testo. Le motivazioni, i dati usati, le metriche di valutazione, e risultati post-valutazione sono presentati e discussi.*

1 Introduction and Motivation

Reasoning about events plays a fundamental role in text understanding. It involves many aspects such as the identification and classification of events, the identification of event participants, the anchoring and ordering of events in time, and their factuality profiling.

In the context of the 2016 EVALITA evaluation campaign, we organized FactA (*Event Factuality Annotation*), the first evaluation exercise for factuality profiling of events in Italian. The task is a follow-up of Minard et al. (2015) presented in the track "Towards EVALITA 2016" at CLiC-it 2015. Factuality profiling is an important component for the interpretation of the events in discourse. Different inferences can be made from events which have not happened (or whose happening is probable) than from those which are described as factual. Many NLP applications such as Question Answering, Summarization, and Textual Entailment, among others, can benefit from the availability of this type of information.

Factuality emerges through the interaction of linguistic markers and constructions and its annotation represents a challenging task. The notion of factuality is strictly related to other research areas throughly explored in NLP, such as subjectivity, belief, hedging and modality (Wiebe et al., 2004; Prabhakaran et al., 2010; Medlock and Briscoe, 2007; Saurı et al., 2006). In this work, we adopted a notion of factuality which corresponds to the committed belief expressed by relevant sources towards the status of an event (Saurí and Pustejovsky, 2012). In particular, the factuality profile of events is expressed by the intersections of two axes: i.) certainty, which expresses a continuum which ranges from absolutely certain to uncertain; and ii.) polarity, which defines a binary distinction: affirmed (or positive) *vs.* negated (or negative).

In recent years, factuality profiling has been the focus of several evaluation exercises and shared tasks, especially for English, both in the newswire domain and in the biomedical domain. To mention the most relevant:

- the BioNLP 2009 Task 3 [1] and BioNLP 2011 Shared [2] Task aimed at recognizing if biomolecular events were affected by speculation or negation;
- the CoNLL 2010 Share Task [3] focused on hedge detection, i.e. identify speculated events, in biomedical texts;
- the ACE Event Detection and Recognition

[1] http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/
[2] http://2011.bionlp-st.org
[3] http://rgai.inf.u-szeged.hu/index.php?lang=en&page=conll2010st

tasks [4] required systems to distinguish between asserted and non-asserted (e.g. hypothetical, desired, and promised) extracted events in news articles;

- the 2012 *SEM Shared Task on Resolving The Scope of Negation [5] focused one of its substasks on the identification of negated, i.e. counterfactual, events;

- the Event Nugget Detection task at TAC KBP 2015 Event Track [6] aimed at assessing the performance of systems in identifying events and their factual, or *realis*, value in news (Mitamura et al., 2015);

- the 2015 [7] and 2016 [8] SemEval Clinical TempEval tasks required systems to assign the factuality value (i.e. attributed modality and polarity) to the extracted events in clinical notes.

Finally recent work, such as the Richer Event Description annotation initiative,[9] has extended the annotation of factuality on temporal relations between pairs of events or pairs of events and temporal expressions as a specific task, independent from the factuality of the events involved, to represent claims about the certainty of the temporal relations themselves.

FactA provides the research community with new benchmark datasets and an evaluation environment to assess system performance concerning the assignment of factuality values to events. The evaluation is structured in two tasks: a Main Task, which focuses on the factuality profile of events in the newswire domain, and a Pilot Task, which addresses the factuality profiling of events expressed in tweets. To better evaluate system performance on factuality profiling and avoid the impact of errors from related subtasks, such as event identification, we restricted the task to the assignment of factuality values. Although as many as 13 teams registered for the task, none of those teams actually submitted any output. Nevertheless, we were able to run an evaluation following the evaluation campaign conditions for one system which was developed by one of the organizers, FactPro.

The remainder of the paper is organized as follows: the evaluation exercise is described in detail in Section 2, while the datasets are presented in Section 3. In Section 4 we describe the evaluation methodology and in Section 5 the results obtained with the FactPro system are illustrated. We conclude the paper in Section 6 with a discussion about the task and future work.

2 Task Description

Following Tonelli et al. (2014) and Minard et al. (2014), in FactA we represent factuality by means of three attributes associated with events,[10] namely *certainty*, *time*, and *polarity*. The FactA task consisted of taking as input a text in which the textual extent of events is given (i.e. gold standard data) and assign to the events the correct values for the three factuality attributes [11] according to the relevant source. In FactA, the relevant source is either the utterer (in direct speech, indirect speech or reported speech) or the author of the news (in all other cases). Systems do not have to provide the overall factuality value (FV): this is computed automatically on the basis of the *certainty*, *time*, and *polarity* attributes (see Section 2.2 for details).

2.1 Factuality Attributes

Certainty. *Certainty* relates to how sure the relevant source is about the mentioned event and admits the following three values: `certain` (e.g. 'rassegnato' in [1]), `non_certain` (e.g. 'usciti' in [2]), and `underspecified` (e.g. 'spiegazioni' in [3]).

1. *Smith ha rassegnato ieri le dimissioni; nomineranno il suo successore entro un mese. ("Smith resigned yesterday; they will appoint his replacement within a month.")*

[4] `http://itl.nist.gov/iad/mig/tests/ace/`

[5] `http://ixa2.si.ehu.es/starsem/index.php\%3Foption=com_content&view=article&id=52&Itemid=60.html`

[6] `http://www.nist.gov/tac/2015/KBP/Event/index.html`

[7] `http://alt.qcri.org/semeval2015/task6/`

[8] `http://alt.qcri.org/semeval2016/task12/`

[9] `https://github.com/timjogorman/RicherEventDescription/blob/master/guidelines.md`

[10] Based on the TimeML specifications (Pustejovsky et al., 2003), the term *event* is used as a cover term for situations that happen or occur, including predicates describing states or circumstances in which something obtains or holds true.

[11] Detailed instruction are reported in the FactA Annotation Guidelines available at `http://facta-evalita2016.fbk.eu/documentation`

2. *Probabilmente i ragazzi sono usciti di casa tra le 20 e le 21. ("The guys went probably out between 8 and 9 p.m.")*

3. *L'Unione Europea ha chiesto "spiegazioni" sulla strage di Beslan. ("The European Union has asked for an explanation about the massacre of Beslan.")*

Time. *Time* specifies the time when an event is reported to have taken place or is going to take place. Its values are `past/present` (for non-future events, e.g. 'capito' in [4]), `future` (for events that will take place, e.g. 'lottare' in [4]) or 'nomineranno' in [1]), and `underspecified` (e.g. 'verifica' in [5]).

4. *I russi hanno capito che devono lottare insieme. ("Russians have understood that they must fight together.")*

5. *Su 542 aziende si hanno i dati definitivi mentre per le altre 38 si è tuttora in fase di verifica. ("They have the final data for 542 companies while for the other 38 it is under validation.")*

Polarity. *Polarity* captures whether an event is affirmed or negated and, consequently, it can be either `positive` (e.g. 'rassegnato' in [1]) or `negative` (e.g. 'nominato' in [6]); when there is not enough information available to detect the *polarity* of an event mention, its value is `underspecified` (e.g. 'scompone' in [7]).

6. *Non ha nominato un amministratore delegato. ("He did not appoint a CEO.")*

7. *Se si scompone il dato sul nero, si vede che il 23% è dovuto a lavoratori residenti in provincia. ("If we analyze the data about the black market labor, we can see that 23% is due to workers resident in the province.")*

Event mentions in texts can be used to refer to events that do not correlate with a real situation in the world (e.g. 'parlare' in [8]). For these event mentions, participant systems are required to leave the value of all three attributes empty.

8. *Guardate, penso che sia prematuro parlare del nuovo preside. ("Well, I think it is too early to talk about the new dean.")*

2.2 Factuality Value

The combination of the *certainty*, *time*, and *polarity* attributes described above determines the factuality value (FV) of an event with respect to the relevant source.

As shown in Table 1, the FV can assume five values: i.) `factual`; ii.) `counterfactual`; iii.) `non-factual`; iv.) `underspecified`; and v.) `no factuality` (no fact). We illustrate in Table 1 the full set of valid combinations of the attribute values and the corresponding FV.

A `factual` value is assigned if an event has the following configuration of attributes:

- *certainty*: `certain`
- *time*: `past/present`
- *polarity*: `positive`

For instance, the event 'rassegnato' [*resigned*] in [1] will qualify as a `factual` event. On the other hand, a change in the *polarity* attribute, i.e. `negative`, will give rise to a `counterfactual` FV, like for instance the event 'nominato' [*appointed*] in [6].

`Non-factual` events depend on the values of the *certainty* and *time* attributes. In particular, a `non-factual` value is assigned if either of the two cases below occur, namely:

- *certainty*: `non_certain`; or
- *time*: `future`

This is the case for the event 'lottare' [*fight*] in [4], where *time* is `future`, or the event 'usciti' [*went out*] in [2] where *certainty* is `non_certain`.

The event FV is `underspecified` if at least one between *certainty* and *time* is `underspecified`, independently of the *polarity* value, like for instance in the case of 'verifica' [*validation*] in [5].

Finally, if the three attributes have no value, FV is `no factuality` (e.g. 'parlare' [*discuss*] in [8]).

3 Dataset Description

We made available an updated version of Fact-Ita Bank (Minard et al., 2014) as training data to participants. This consists of 169 documents selected from the Ita-TimeBank (Caselli et al., 2011) and

[12]The number of tokens for the pilot test is computed after the tokenizaion, i.e. the hashtags and aliases can be split in more than one token and the emoji are composed by several tokens.

Certainty	Time	Polarity	FV
certain	past/pres.	positive	factual
certain	past/pres.	negative	counterfact.
non_cert.	*any value*	*any value*	non-fact.
any value	future	*any value*	non-fact.
certain	undersp.	*any value*	underspec.
undersp.	past/pres.	*any value*	underspec.
undersp.	undersp.	*any value*	underspec.
-	-	-	no fact.

Table 1: Possible combinations of factuality attributes.

first released for the EVENTI task at EVALITA 2014.[13] Fact-Ita Bank contains annotations for 6,958 events (see Table 2 for more details) and is distributed with a CC-BY-NC license.[14]

As test data for the Main Task we selected the Italian section of the NewsReader MEANTIME corpus (Minard et al., 2016), a corpus of 120 Wikinews articles annotated at multiple levels. The Italian section is called WItaC, the NewsReader Wikinews Italian Corpus (Speranza and Minard, 2015) , and consists of 15,676 tokens (see Table 2).

As test data for the Pilot Task we annotated 301 tweets with event factuality, representing a subsection of the test set of the EVALITA 2016 SENTIPOLC task (Barbieri et al., 2016) (see Table 2).

Training and test data, both for the Main and the Pilot Tasks, are in the CAT (Content Annotation Tool) (Bartalesi Lenzi et al., 2012) labelled format. This is an XML-based stand-off format where different annotation layers are stored in separate document sections and are related to each other and to source data through pointers.

4 Evaluation

Participation in the task consisted of providing only the values for the three factuality attributes (*certainty*, *time*, *polarity*), while the FV score was to be computed through the FactA scorer on top of these values.

The evaluation is based on the micro-average F1 score of the FVs, which is equivalent to the accuracy in this task as all events should receive a FV (i.e. the total numbers of False Positives and False Negatives over the classes are equal). In addition to this, an evaluation of the performance of the systems on the single attributes (using micro-average F1 score, equivalent to the accuracy) will be provided as well. We consider this type of evaluation to be more informative than the one based on the single FV because it will provide evidence of systems' ability to identify the motivations for the assignment of a certain factuality value. To clarify this point, consider the case of an event with FV `non-factual` (certainty `non_certain`, time `past/present` and polarity `positive`). A system might correctly identify that the FV of the event is `non-factual` because certainty is `non_certain`, or erroneously identify that time is `future`.

5 System Results

Unfortunately no participants took part in the FactA task. However, we managed to run an evaluation test with a system for event factuality annotation in Italian, FactPro, developed by one of the organizers and respecting the evaluation campaign conditions. The system was evaluated against both gold standard, i.e. the Main and Pilot tasks. In this section we describe this system and the results obtained on the FactA task.

5.1 FactPro module

FactPro is a module of the TextPro NLP pipeline [15] (Pianta et al., 2008). It has been developed by Anne-Lyse Minard in collaboration with Federico Nanni as part of an internship.

Event Factuality annotation is performed in FactPro in three steps: (1) detection of the polarity of an event, (2) identification of the certainty of an event and (3) identification of the semantic time. These three steps are based on a machine learning approach, using Support Vector Machines algorithm, and are taken as text chunking tasks in which events have to be classified in different classes. For each step a multi-class classification model is built using the text chunker Yamcha (Kudo and Matsumoto, 2003).

FactPro requires the following pre-processes: sentence splitting, tokenization, morphological analysis, lemmatization, PoS tagging, chunking, and event detection and classification. As the data provided for FactA consist of texts already split into sentences, tokenized and annotated with events, the steps of sentence splitting, tokenization and event

[13] `https://sites.google.com/site/eventievalita2014/home`

[14] `http://hlt-nlp.fbk.eu/technologies/fact-ita-bank`

[15] `http://textpro.fbk.eu`

	training set (main) Fact-Ita Bank	**test set (main)** MEANTIME	**test set (pilot)** (tweets)
tokens[12]	65,053	15,676	4,920
sentences	2,723	597	301
events	6,958	1,450	475
certainty			
certain	5,887	1,246	326
non certain	813	133	53
underspecified	204	53	43
time			
past/present	5,289	1,026	263
future	1,560	318	113
underspecified	55	88	46
polarity			
positive	6,474	1,363	381
negative	378	45	27
underspecified	52	24	14
FV			
factual	4,831	978	225
counterfactual	262	32	15
non-factual	1,700	327	126
underspecified	111	95	56
no_factuality	54	18	53

Table 2: Corpora statistics

detection and classification are not performed for these experiments.

Each classifier makes use of different features: lexical, syntactic and semantic. They are described in the remainder of the section. For the detection of polarity and certainty, FactPro makes use of trigger lists which have been built manually using the training corpus.

- Polarity features:
  - For all tokens: token's lemma, PoS tags, whether it is a polarity trigger (list manually built);
  - If the token is part of an event: presence of polarity triggers before it, their number, the distance to the closest trigger, and whether the event is part of a conditional construction;
  - The polarity value tagged by the classifier for the two preceding tokens.
- Certainty features:
  - For all tokens: token's lemma, flat constituent (noun phrase or verbal phrase), whether it is a modal verb, whether it is a certainty trigger (list manually built);
  - If the token is part of an event: the event class (It-TimeML classes), presence of a modal before and its value, and whether the event is part of a conditional construction;
  - The certainty value tagged by the classifier for the two preceding tokens.
- Time features:
  - For all tokens: token's lemma and whether it is a preposition;
  - If the token is part of an event: tense and mood of the verb before, presence of a preposition before, event's polarity and certainty;
  - If the token is a verb: its tense and mood;
  - The time value tagged by the classifier for the three preceding tokens.

Each token is represented using these features as well as some of the features of the previous tokens and of the following ones. We have defined the set of features used by each classifier performing several evaluations on a subsection of the Fact-Ita Bank corpus.

task	system	polarity	certainty	time	3 attributes	Factuality Value
main	baseline	0.94	0.86	0.71	0.67	0.67
main	FactPro	0.92	0.83	0.74	0.69	**0.72**
pilot	baseline	0.80	0.69	0.55	0.47	0.47
pilot	FactPro	0.79	0.66	0.60	0.51	**0.56**

Table 3: Evaluation of FactPro against the baseline (accuracy)

5.2 Results

Table 3 shows the results of FactPro for the two tasks of FactA against a baseline. The baseline system annotates all events as `factual` (the predominant class), i.e. being `certain`, `positive` and `past/present`. The performance of FactPro on the Main Task is 0.72 when evaluating the Factuality Value assignment and 0.69 on the combination of the three attributes, and on the Pilot Task 0.56 and 0.51 respectively. On these two tasks FactPro performs better than the baseline. It has to be noted that we ran FactPro on the pilot test set without any adaptation of the different processes.

In Table 4 we present the F1-score obtained for each value of the three attributes as well for each Factuality Value. We can observe that FactPro does not perform well on the identification of the `underspecified` values and on the detection of events that do not have a factuality value (`no fact`).

5.3 Error Analysis of FactPro

We can observe from Table 3 that FactPro performs better for the detection of polarity and certainty than for the identification of time. One reason is the predominance of one value for the polarity and certainty attributes, and of two values for time. For example, in the training corpus, 94% of the events have a polarity `positive` and 86% are `certain`, whereas 71% of the events are `past/present` and 22% are `future`.

An extensive error analysis on the output of the systems for the three attributes was conducted. As for the polarity attribute, the error analysis showed that the system's failure to detect negated events is not mainly due to a sparseness of negated events in the training data, but it mainly concerns the negation scope, whereas when the system missed a negative event it was mainly due to the incompleteness of the trigger lists (e.g. *mancata* in *dopo la mancata approvazione* is a good trigger for polarity `negative` but it is absent from the trigger list).

The detection of `non_certain` events works well when the event is preceded by a verb at the conditional and when it is part of an infinitive clause introduced by *per*. However when the uncertainty of an event is expressed by the semantics of previous words (e.g. *salvataggio* in *il piano di salvataggio*) the system makes errors.

With respect to the annotation of `future` events, the observations are similar to those for `non_certain` events. Indeed, future events are well recognized by the system when they are part of an infinitive clause introduced by the preposition *per* as well as when their tense is future.

Finally, we observed that FactPro makes a lot of errors when the annotation of the factuality of nominal events is concerned. In the Main Task it correctly identified the FV of 81% of the verbal events and only 61% of the nominal events.

6 Conclusion and Future Work

The lack of participants in the task limits the discussion of the results to the in-house developed system. The main reason for the lack of participation to FactA, according to the outcome of a questionnaire organized by the 2016 EVALITA chairs, was that the participants gave priority to other EVALITA tasks. However, FactA achieves two main results: i.) setting state-of-the-art results for the factuality profiling of events in two text types in Italian, namely news articles and tweets; and ii.) making available to the community a new benchmark corpus and standardized evaluation environment for comparing systems' performance and facilitating replicability of results.

The test data used for the Main Task consists of the Italian section of the MEANTIME corpus (Minard et al., 2016). MEANTIME contains the same documents aligned in English, Italian, Spanish and Dutch, thus making available a multilingual environment for cross-language evaluation of the factuality profiling of events. Furthermore, within the NewsReader project, a module for event factuality annotation has been implemented and evaluated against the English section of the MEANTIME

	polarity			certainty			time		
task	pos.	neg.	undersp.	cert.	non_cert.	undersp.	past/pres.	future	undersp.
main	0.96	0.68	0.00	0.91	0.42	0.10	0.84	0.54	0.00
pilot	0.88	0.69	0.00	0.80	0.35	0.18	0.73	0.50	0.00

	FV				
task	factual	counterfact.	non-fact.	undersp.	no fact.
main	0.83	0.62	0.55	0.02	0.00
pilot	0.72	0.39	0.50	0.03	0.29

Table 4: FactPro results on the single attribute and on the different factuality value (F1 score)

corpus (Agerri et al., 2015). The evaluation was performed in a different way than in FactA, in particular no gold events were provided as input to the system, so the evaluation of factuality was done only for the events correctly identified by the event detection module. The system obtained an accuracy of 0.88, 0.86 and 0.59 for polarity, certainty, and time, respectively.

The Pilot task was aimed at evaluating how well systems built for standard language perform on social media texts, and at making available a set of tweets annotated with event mentions (following TimeML definition of events) and their factuality value. The pilot data are shared between three other tasks of EVALITA 2016 (PoSTWITA, NEEL-IT and SENTIPOLC), which contributed to the creation of a richly annotated corpus of tweets to be used for future cross-fertilization tasks. Finally, the annotation of tweets raised new issues for factuality annotation because tweets contain a lot of imperatives and interrogatives that are generally absent from news and for which the factuality status is not obvious (e.g. *Ordini solo quello che ti serve*).

The results obtained by FactPro, as reported in Table 3 and Table 4, show that i.) the system is able to predict with pretty high accuracy the FV on events in the news domain and with a lower but good score the factuality of events in tweets; ii.) the difference in performance between the news and tweet text types suggest that specific training set data may be required to address the peculiarities of tweets' language; iii.) the F1 scores for the certainty, polarity and time attributes clearly indicate areas of improvements and also contribute to a better understanding of the system's results; iv.) the F1 scores on the FV suggest that extending the training data with tweets could also benefit the identification of values which are not frequent in the news domain, such as `no_fact`.

Future work will aim at re-running the task from raw text and developing specific modules for the factuality of events according to the text types where they occur. Finally, we will plan to run a cross-fertilization task concerning temporal ordering and anchoring of events and factauality profiling.

Acknowledgments

This work has been partially supported by the EU NewsReader Project (FP7-ICT-2011-8 grant 316404) and the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3).

References

Rodrigo Agerri, Itziar Aldabe, Zuhaitz Beloki, Egoitz Laparra, German Rigau, Aitor Soroa, Marieke van Erp, Antske Fokkens, Filip Ilievski, Ruben Izquierdo, Roser Morante, Chantal van Son, Piek Vossen, and Anne-Lyse Minard. 2015. Event Detection, version 3. Technical Report D4-2-3, VU Amsterdam. `http://kyoto.let.vu.nl/newsreader_deliverables/NWR-D4-2-3.pdf`.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Linguistic Annotation Workshop*, pages 143–151.

Taku Kudo and Yuji Matsumoto. 2003. Fast Methods for Kernel-based Text Analysis. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 24–31, Stroudsburg, PA, USA.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *ACL*, volume 2007, pages 992–999. Citeseer.

Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. 2014. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. In *Proceedings of CLiC-it 2014, First Italian Conference on Computational Linguistic*.

Anne-Lyse Minard, Manuela Speranza, Rachele Sprugnoli, and Tommaso Caselli. 2015. FacTA: Evaluation of Event Factuality and Temporal Anchoring. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoa Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may.

Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event Nugget Annotation: Processes and Issues. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 66–76.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1014–1022. Association for Computational Linguistics.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, pages 28–34.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Roser Saurı, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of 19th International FLAIRS Conference*.

Manuela Speranza and Anne-Lyse Minard. 2015. Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC). In *Proceedings of CLiC-it 2015, Second Italian Conference on Computational Linguistic*.

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. 2014. NewsReader Guidelines for Annotation at Document Level. Technical Report NWR2014-2-2, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf`.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.