# Parsing Events: a New Perspective on Old Challenges

**Rachele Sprugnoli[1], Felice Dell'Orletta[2], Tommaso Caselli[3], Simonetta Montemagni[2], Cristina Bosco[4]**
[1]Fondazione Bruno Kessler - Università di Trento
[2]Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR, Pisa
[3]VU Amsterdam, [4]Dipartimento di Informatica - Università di Torino
sprugnoli@fbk.eu, felice.dellorletta@ilc.cnr.it,
t.caselli@vu.nl
simonetta.montemagni@ilc.cnr.it, bosco@di.unito.it

## Abstract

**English.** The paper proposes a new evaluation exercise, meant to shed light on the syntax-semantics interface for the analysis of written Italian and resulting from the combination of the EVALITA 2014 dependency parsing and event extraction tasks. It aims at investigating the cross-fertilization of tasks, generating a new resource combining dependency and event annotations, and devising metrics able to evaluate the applicative impact of the achieved results.

**Italiano.** *L'articolo propone un innovativo esercizio di valutazione focalizzato sull'interfaccia sintassi-semantica per l'analisi dell'italiano scritto che combina i task di EVALITA 2014 su parsing a dipendenze ed estrazione di eventi. Il suo contributo consiste nell'approfondire la combinazione di task che spaziano tra diversi livelli di analisi, nello sviluppo di nuove risorse con annotazione a dipendenze e basata su eventi, e nella proposta di metriche che valutino l'impatto applicativo dei risultati ottenuti.*

## 1 Introduction

Since the '90s, evaluation campaigns organized worldwide have offered to the computational linguistics community the invaluable opportunity of developing, comparing and improving state-of-the-art technologies in a variety of NLP tasks. ACE[1], MUC[2], CoNLL[3] and SemEval[4] are probably the best-known series of evaluation campaigns that covered syntactic and semantic tasks for English as well as for other languages (e.g. Spanish, Arabic, Chinese). For Italian, EVALITA campaigns[5] have been organized since 2007 around a set of evaluation exercises related to the automatic analysis of both written text and speech.

Over the years, many challenging tasks have been proposed with the aim of advancing state-of-the-art technologies in different NLP areas: to mention only a few, dependency parsing (Nivre et al., 2007), (Bosco and Mazzei, 2013), textual entailment (Bos et al., 2009), frame labeling (Basili et al., 2013) and cross-document event ordering (Minard et al., 2015), all requiring cutting-edge methods and techniques as well as innovative approaches.

Following the fact that, in recent years, research is moving from the analysis of grammatical structure to sentence semantics, the attention in evaluation campaigns is shifting towards more complex tasks, combining syntactic parsing with semantically-oriented analysis. The interest of composite and articulated tasks built by combining basic tasks also lies at the applicative level, since Information Extraction architectures can realistically be seen as integrating components which carry out distinct basic tasks.

Starting from the analysis of the results achieved for individual tasks in EVALITA 2014 and illustrated in Attardi et al. (2015), this paper represents a first attempt of designing a complex shared task for the next EVALITA edition, resulting from the combination of the dependency parsing and event extraction tasks for the analysis of Italian texts. Such a complex task is expected to shed new light onto old challenges by: a.) investigating whether and how the cross-fertilization of tasks can make the evaluation campaign more application-oriented, while also improving individual task results; b.) generating a new resource combining dependency

[1]http://www.itl.nist.gov/iad/mig/tests/ace/
[2]http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
[3]http://ifarm.nl/signll/conll/
[4]http://aclweb.org/aclwiki/index.php?title=SemEval_Portal
[5]http://www.evalita.it

and event annotation; and, c.) devising evaluation metrics more oriented towards the assessment of the applicative impact of the achieved results.

## 2 Motivation and Background

In recent years, syntactic and semantic dependency parsing have seen great advances thanks to the large consensus on representation formats and to a series of successful evaluation exercises at CoNLL (Surdeanu et al., 2008; Hajič et al., 2009) and SemEval (Oepen et al., 2014; Oepen et al., 2015). However, access to the content, or meaning, of a text has not reached fully satisfactory levels yet. Current developments of data-driven models of parsing show that the recovery of the full meaning of text requires simultaneous analysis of both its grammar and its semantics (Henderson et al., 2013), whose interaction is still not well understood and varies cross-linguistically.

Since the CoNLL 2008 shared task (Surdeanu et al., 2008) much research has focused on the development of systems able either to jointly perform syntactic and semantic dependency tasks or to tackle them independently by means of pipelines of NLP modules specialized in the various subtasks (first full syntactic parsing and then semantic parsing). Insights on the linguistic relatedness of the two tasks derived from the comparison of joint and disjoint learning systems results. Another example is the SemEval 2010 "Task 12: Parser Evaluation using Textual Entailments (Yuret et al., 2010)" (PETE), aimed at recognizing textual entailment based on syntactic information only and whose results highlighted semantically relevant differences emerging from syntax. The evaluation exercise is closer to an extrinsic evaluation of syntactic parsing by focusing on semantically relevant differences.

At EVALITA 2014, two evaluation exercises for the analysis of written text, Dependency Parsing (Bosco et al., 2014) and EVENTI (Caselli et al., 2014), have provided separate evaluations of these two levels of analysis: syntax and semantics, respectively. The relation between the two levels of analysis was investigated in the Dependency Parsing task by setting up a semantically-oriented evaluation assessing the ability of participant systems to produce suitable and accurate output for Information Extraction. Based on measures such as Precision, Recall and F1, this evaluation has been carried out against a subset of 19 semantically-loaded dependency relations (e.g. subject, direct object, adjectival complement and temporal modifier among others). On the other hand, in the EVENTI exercise, syntactic information was considered to play a relevant role for at least two of the subtasks: event detection and classification (subtask B) and temporal relation identification and classification (subtask C).

Dependency parsing is now a key step of analysis from which higher-level tasks (e.g. semantic relations, textual entailment, temporal processing) can definitely benefit. Event Extraction is a high-level semantic task which is strictly connected to morphology and syntax both for the identification of the event mentions and for their classification. Event Extraction differs from standard semantic parsing as not all event mentions have semantic dependencies and it involves a wider range of linguistic realizations (such as verbs, nouns, adjectives, and prepositional phrases) some of which have not been taken into account so far in standard semantic parsing tasks. Despite the recognized influence of one level of analysis on the other, no systematic bi-directional analysis has been conducted so far. To gain more insight on the syntax-semantics interface more focused and complex evaluation exercises need to be setup and run.

In this paper we propose a new evaluation exercise, named "Parsing Events", which aims at shedding new light on the syntax-semantics interface in the analysis of Italian written texts by investigating whether and to what extent syntactic information helps improving the identification and classification of events, and conversely whether and to what extent semantic information, event mentions and classes, improve the identification and classification of dependency relations.

## 3 Task Description

Parsing Events will qualify as a new evaluation exercise for promoting research in Information Extraction and access to the text meaning for Italian. The exercise, which will start from previous research and datasets for Dependency Parsing and Temporal Processing of Italian, aims at opening a new perspective for what concerns the evaluation of systems to be carried out both at a high level, targeting complex Information Extraction architectures, and at a low level, as single components. The Parsing Events exercise will be thus articulated as follows: a main task, joint dependency parsing and event extraction, and two subtasks, dependency

parsing and event extraction, respectively.

**Main task - Joint Dependency Parsing and Event Extraction**: The main task will test systems for Dependency Parsing and Event Extraction. Systems have to determine dependency relations based on the ISDT[6] (Bosco et al., 2013) scheme and identify all event mentions as specified in the EVENTI annotation guidelines (Caselli and Sprugnoli, 2014). This will imply to identify the event mentions and fill the values of target attributes. To better evaluate the influence of syntactic information in Event Extraction, the set of event attributes which will be evaluated will be extended to include CLASS, TENSE, ASPECT, VFORM, MOOD and POLARITY. Participants will be given annotated data with both syntactic and event annotations for training. Ranking will be performed on the F1 score of a new evaluation measure based on Precision and Recall for event class and dependency relation.

**Subtask A - Dependency parsing** The subtask on Dependency Parsing will be organized as a classical dependency parsing task, where the performance of different parsers can be compared on the basis of the same set of test data provided by the organizers. The main novelty of this task with respect to the traditional dependency parsing task organized in previous EVALITA campaigns is that available information will also include event-related information.

**Subtask B - Event extraction** The Event Extraction subtask will be structured as the Subtask B of the EVENTI 2014 evaluation (Caselli et al., 2014). Participants will be asked to identify all event mentions according to the EVENTI annotation guidelines. The set of event attributes which will be evaluated is extended as described in the Main Task. The main innovation with respect to the original task is that participants will be provided with dependency parsing data both in training and test. Systems will be ranked according to the attribute CLASS F1 score.

### 3.1 Annotation and Data Format

In the spirit of re-using available datasets, the annotation efforts for the Parsing Events task will be mainly devoted to the creation of a new test set, called Platinum data, which will contain manual annotation for both dependency parsing and events. The size of the Platinum data will be around 10k-20k tokens. The annotation of the dataset will be conducted by applying the ISDT guidelines for the dependency parsing information and the EVENTI guidelines for events. An innovative aspect of the Platinum data concerns the text genres. To provide a more reliable evaluation, the Platinum data will consist of newspaper articles and biographies from Wikipedia[7].

The training data (Gold data) will be based on the EVENTI and the Dependency Parsing data. A subset of 27,597 tokens between the two datasets perfectly overlaps, thus making already available Gold annotations. Given that the focus of the evaluation exercise is on the reciprocal influence of the two basic tasks, we will provide the missing annotations on the remaining parts (i.e. 102,682 tokens for the EVENTI dataset and 160,398 tokens for the Dependency Parsing dataset) by means of automatically generated annotation, i.e. Silver data. Silver data have already been successfully used to extend the size of training data in previous evaluation exercises (e.g. TempEval-3). Furthermore, we plan to extend the set of overlapping Gold data by manual revision.

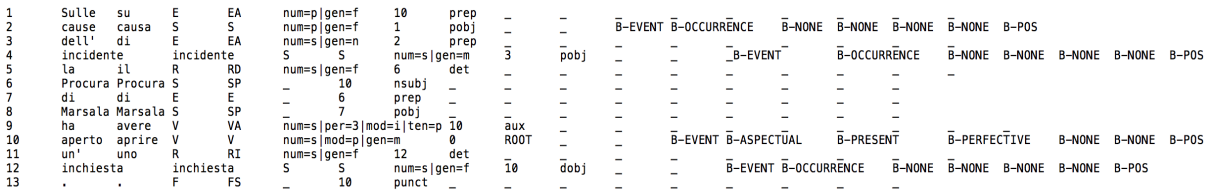
Training data will be distributed in a unified representation format compliant with the CoNLL-X specifications (Buchholz and Marsi, 2006) and extended for the encoding of event information which will be annotated in terms of standard IOB representation as exemplified in Figure 1 (the example is taken from the overlapping portion of the training data of the two task at EVALITA 2014). Event annotation (last seven columns) is concerned with the following information types: event extent, class, tense, aspect, vform, mood and polarity.

The test set for the main task will be distributed in the same format of the training dataset providing participants with pre-tokenized, POS-tagged and lemmatized data. This distribution format will be adopted also for the two subtasks. In addition to the information regarding tokens, POS tags and lemmas, Gold data for events will be available for the dependency parsing subtask, while Gold data for dependency parsing will be available for the event extraction subtask.

Systems will be required to produce a tab-delimited file. Systems participating to the main task will provide in output the extended CoNLL-X format including the information for the event an-

[6] http://medialab.di.unipi.it/wiki/ISDT

[7] The biographical data are part of the multilingual parallel section (Italian / English) of TUT (ParTUT http://www.di.unito.it/~tutreeb/partut.html).



| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Sulle | su | E | EA | num=p\|gen=f | 10 | prep | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ |
| 2 | cause | causa | S | S | num=p\|gen=f | 1 | pobj | _ | _ | B-EVENT | B-OCCURRENCE | B-NONE | B-NONE | B-NONE | B-NONE | B-POS | | | |
| 3 | dell' | di | E | EA | num=s\|gen=n | 2 | prep | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ |
| 4 | incidente | | incidente | | S | S | num=s\|gen=m | 3 | pobj | _ | _ | B-EVENT | | B-OCCURRENCE | | B-NONE | B-NONE | B-NONE | B-NONE | B-POS |
| 5 | la | il | R | RD | num=s\|gen=f | 6 | det | _ | _ | _ | _ | _ | _ | _ | _ | _ | | | | |
| 6 | Procura | Procura | S | SP | _ | 10 | nsubj | _ | _ | _ | _ | _ | _ | _ | _ | _ | | | | |
| 7 | di | di | E | E | _ | 6 | prep | _ | _ | _ | _ | _ | _ | _ | _ | _ | | | | |
| 8 | Marsala | Marsala | S | SP | _ | 7 | pobj | _ | _ | _ | _ | _ | _ | _ | _ | _ | | | | |
| 9 | ha | avere | V | VA | num=s\|per=3\|mod=i\|ten=p | 10 | aux | _ | _ | _ | _ | _ | _ | _ | _ | _ | | | | |
| 10 | aperto | aprire | V | V | num=s\|mod=p\|gen=m | 0 | ROOT | _ | _ | B-EVENT | B-ASPECTUAL | | B-PRESENT | | B-PERFECTIVE | | B-NONE | B-NONE | B-POS |
| 11 | un' | uno | R | RI | num=s\|gen=f | 12 | det | _ | _ | _ | _ | _ | _ | _ | _ | _ | | | | |
| 12 | inchiesta | | inchiesta | | S | S | num=s\|gen=f | 10 | dobj | _ | _ | B-EVENT | B-OCCURRENCE | | B-NONE | B-NONE | B-NONE | B-NONE | B-POS |
| 13 | . | . | F | FS | _ | 10 | punct | _ | _ | _ | _ | _ | _ | _ | _ | _ | | | | |

Figure 1: Example of a complete annotated sentence with syntactic and event information.

notation as shown in Figure 1. Systems taking part to the individual subtasks will provide in output the relevant fields: head token id, and the dependency linking the token under description to its head, for the dependency parsing subtask; the event extent and associated attributes for the event extraction subtask.

## 4 Evaluation and Discussion

Evaluation of systems is not a trivial issue. For the evaluation of participating systems we foresee at the moment different evaluation metrics for each task, described below.

**Main Task**: The main task aims at evaluating the bi-directional influence of syntactic and semantic information. We are then proposing a hybrid measure which takes into account the correctness of the event class and that of the dependency label. We propose the following definitions of Precision, Recall, and F1:

- Precision: the ratio between the tokens with correct event class and labeled dependency from the system, $tp_i$, and all tokens marked as event by the system ($tp_i$ and $fp_i$): $\frac{tp_i}{tp_i+fp_i}$;
- Recall: the tokens with correct event class and labeled dependency from the system, $tp_i$, and the number of positive examples in the Gold data ($tp_i$ plus false negatives $fn_i$) : $\frac{tp_i}{tp_i+fn_i}$
- F1: the mean of Precision and Recall calculated as follows: $\frac{2 Precision Recall}{Precision+Recall}$

**Subtask A**: Similarly to the dependency parsing task presented in EVALITA 2014, in addition to the standard accuracy dependency parsing measures, i.e. Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS), we will provide an alternative and semantically-oriented metric to assess the ability of the parsers to produce reliable and accurate output for Information Extraction applications. As in EVALITA 2014, we will select a set of dependency relations and for these relations the parser accuracy will be evaluated using Precision, the ratio of correct relations extracted over the total of extracted relations; Recall, the ratio of correct relations extracted over the relations to be found (according to the gold standard); and F-Measure. Differently from EVALITA 2014, for this semantically-oriented evaluation we will focus on dependency relations involved in the syntax of event structures.

**Subtask B**: Following the EVENTI evaluation exercise, the Event Extraction subtask will be evaluated by applying the adapted TempEval-3 scorer (UzZaman et al., 2013; Caselli et al., 2014). We will evaluate i.) the number of the elements correctly identified and if their extension is correct, and ii.) the attribute values correctly identified. As for the first aspect, we will apply standard Precision, Recall and F1 scores. Strict and relaxed (or partial) match will be taken into account. On the other hand, attribute evaluation will be computed by means of the attribute F1 score (UzZaman et al., 2013), which measures how well a system identified the element and corresponding attributes' values.

For the evaluation of subtask results, participants will be asked to submit different runs, carried out with and without the information from the other subtask: i.e. Dependency Parsing will be carried out with and without event information, and Event Extraction will be carried out with and without dependency information. This type of contrastive evaluation highlights one of the main novelties of the proposed complex task, which is not only aimed at assessing the performance of participating systems and ranking achieved results, but also at investigating impact and role of different types of information on each task depending on the adopted algorithm. A shared task organized along these lines thus creates the prerequisites for a more accurate error analysis and will possibly open up new directions of research in tackling old challenges.

## References


Giuseppe Attardi, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. 2015. State of the art language technologies for Italian: The EVALITA 2014 perspective. *Intelligenza Artificiale*, 9(1):43–61.

Roberto Basili, Diego De Cao, Alessandro Lenci, Alessandro Moschitti, and Giulia Venturi. 2013. EVALITA 2011: The frame labeling over Italian texts task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, Lecture Notes in Computer Science, pages 195–204. Springer Berlin Heidelberg.

Johan Bos, Fabio Massimo Zanzotto, and Marco Pennacchiotti. 2009. Textual entailment at EVALITA 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

Cristina Bosco and Alessandro Mazzei. 2013. The EVALITA dependency parsing task: from 2007 to 2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, Lecture Notes in Computer Science, pages 1–12. Springer Berlin Heidelberg.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.

Cristina Bosco, Felice DellOrletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. In *Proceedings of the Fourth International Workshop EVALITA 2014*, pages 1–8.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Tommaso Caselli and Rachele Sprugnoli. 2014. EVENTI Annotation Guidelines for Italian v.1.0. Technical report, FBK and TrentoRISE.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI. EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*, pages 27–34.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

James Henderson, Paola Merlo, Ivan Titov, and Gabriele Musillo. 2013. Multilingual joint parsing of syntactic and semantic dependencies with a latent variable model. *Computational Linguistics*, 39(4):949–998.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Rubén Urizar. 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan T. McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 915–932.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Yi Zhang. 2014. Semeval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova, Dan Flickinger, Jan Hajic, and Zdenka Uresova. 2015. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado, June. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

Nashaud UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9. Association for Computational Linguistics, Atlanta, Georgia, USA.

Deniz Yuret, Aydin Han, and Zehra Turgut. 2010. Semeval-2010 task 12: Parser evaluation using textual entailments. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 51–56. Association for Computational Linguistics.