Natural Language Engineering: page 1 of 22. © Cambridge University Press 2016
doi:10.1017/S1351324916000292

*One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective**

R. SPRUGNOLI[1,2] and S. TONELLI[1]

[1]*Fondazione Bruno Kessler, Via Sommarive 18, Trento, Italy*
e-mails: sprugnoli@fbk.eu, satonelli@fbk.eu
[2]*Università di Trento, Via Sommarive 9, Trento, Italy*

(*Received 28 August 2015; revised 12 September 2016; accepted 13 September 2016*)

Abstract

We present an overview of event definition and processing spanning 25 years of research in NLP. We first provide linguistic background to the notion of event, and then present past attempts to formalize this concept in annotation standards to foster the development of benchmarks for event extraction systems. This ranges from MUC-3 in 1991 to the Time and Space Track challenge at SemEval 2015. Besides, we shed light on other disciplines in which the notion of event plays a crucial role, with a focus on the historical domain. Our goal is to provide a comprehensive study on event definitions and investigate which potential past efforts in the NLP community may have in a different research domain. We present the results of a questionnaire, where the notion of event for historians is put in relation to the NLP perspective.

1 Introduction

In the last 25 years, several systems performing event extraction have been presented within the NLP community. Diverse approaches aimed at building timelines from large document collections have been implemented, and technologies to support automatic storytelling have become a relevant research topic in the AI community (Ashish *et al.* 2006). Event processing has been addressed from a variety of perspectives, from data visualization to knowledge representation and modelling. However, the notion of event has been revised several times and often tailored to the task of interest, so that a number of different definitions of event has been introduced since the first MUC evaluation campaign. Furthermore, the notion of event has been studied also in other disciplines, such as philosophy, cognitive science and history, which the NLP community has hardly taken into account.

A further distinction concerns two different research areas within NLP: In the field of Topic Detection and Tracking, the identification of events is assimilated to

*The title is inspired by the novel 'One, No One and One Hundred Thousand' written by the Nobel prize winner Luigi Pirandello.

the identification of topics within a stream of texts and the clustering of documents by topic.[1] Instead, in the field of Information Extraction (IE), the aim is to extract events expressed by words or phrases in a text. In this paper, we focus mainly on the latter perspective, since it has led to several standardization proposals and evaluation campaigns, and to the creation of a wide community of researchers working at Temporal Information Processing tasks. However, we are aware that Topic Detection and Tracking is going to attract more and more attention, because it is particularly suitable to perform coarse-grained event detection on large streams of documents, for instance on social media data.

Temporal Information Processing is a task that aims at automatically detecting and interpreting events (e.g., *to live/the war*), temporal expressions (e.g., *20/05/2015/this summer*) and temporal relations within texts (e.g., in *Waters recede before a tsunami* the event *recede* happens BEFORE the event *tsunami*). Although event identification and processing may appear an easier task than the classification of temporal relations and expressions, which are often vague or implicit in natural language, this is still very challenging due to the ambiguous nature of the concept of event. The term 'event' itself has many readings: some authors use it to refer only to dynamic actions, others to refer also to static situations (Sasse 2002). This terminological confusion mirrors the inherent complexity of the concept of event: In fact, an event may designate both an ontological and a linguistic category. However, between the ontological level and the linguistic one, there is no one-to-one mapping because the same event may be expressed using various types of linguistic elements. As a matter of fact, even if verbs prototypically denote events whereas nominals denote objects, this distinction is not clear-cut in natural language (Hagège 1996). In particular, nominals exhibit a strong semantic ambiguity due to polysemy, showing alternations between eventive and non-eventive readings (Pustejovsky 2005): for example, *administration* denotes an event in *spending grew during his administration* and a group in *this administration is doing well*. A variety of terminology is in use also in event classification. The best-known classification of events is the one proposed by Vendler (1967), who distinguishes between states (non-dynamic situations persisting over a period of time and without an endpoint, e.g., *believe*), activities (open-ended dynamic processes, e.g., *walk*), accomplishments (processes with a natural endpoint and an intrinsic duration, e.g., *build a house*) and achievements (almost instantaneous events with an endpoint, e.g., *find*). The Generative Lexicon theory revisits Vendler's classification introducing a three-way taxonomy of event types including states, processes and transitions. In the latter category, accomplishments and achievements are collapsed (Pustejovsky 1991). Moreover, in the linguistic literature, all types of actions, states and processes often fall under the cover term 'eventualities', coined by Bach (1986) in his work on the algebra of events.

With this survey, we aim at providing an overview of the way events have been defined in IE (Section 2), with a focus on the different evaluation campaigns

[1] According to the LDC annotation guidelines of the TDT task, 'a topic is defined as an event or activity, along with all directly related events and activities', see `https://catalog.ldc.upenn.edu/docs/LDC2005S11/tdt4guidelines_v1_5.pdf`

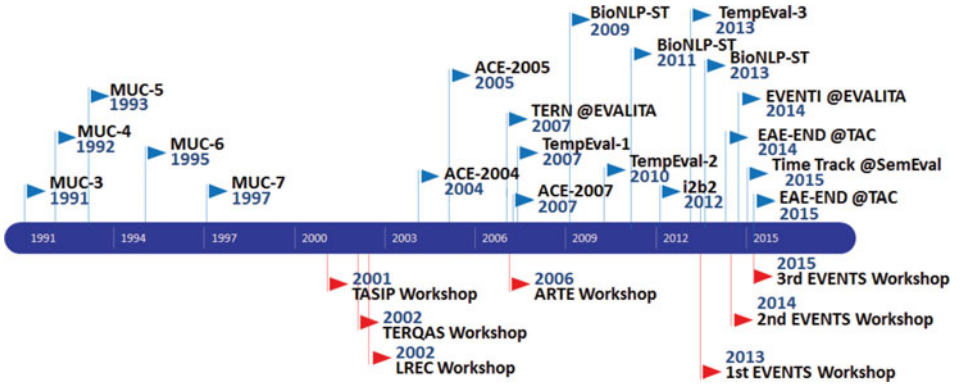

Fig. 1. (Colour online) Timeline of evaluation campaigns (above) and workshops (below) in the field of event detection and processing. The Time & Space Track @Semeval 2015 includes the TimeLine, QA TempEval and Clinical TempEval tasks.

organized over the years (Section 2.2). We also account for multilingual event processing, presenting tasks and corpora that cover languages other than English, and for new domains involved in recent event definition efforts (Section 2.3). Finally, we present a case study in Section 3, taking the perspective of history scholars, i.e., researchers from another area that typically deal with events in their daily activity. We try to address the following questions: Was all the work devoted to event processing with IE techniques useful to serve real historical investigation? Were the various definitions of events provided over the years compatible with research practices adopted in other communities? How should events be defined to be processable with NLP tools but also to comply with historical research? We shed light on such questions by means of an online questionnaire, in which historians were involved in an 'event definition' exercise. The outcome of this study highlights the difficulties in shifting from a linguistic-driven perspective to the historical one, where a more abstract conception of events is prevalent.

2 The IE perspective on events

Starting in 1991, several evaluation campaigns and workshops devoted to various aspects of temporal information processing and in particular to the analysis of the notion of event have been organized and have fostered the creation of a research community around event detection and processing. The timeline in Figure 1, built by collecting information from websites and proceedings, summarizes the history of workshops, in the lower part, and evaluation campaigns, in the upper part, related to temporal processing and organized starting from MUC-3.[2]

We describe them in detail in the following subsections.

[2] An interactive and constantly updated version of the timeline is available online: `http://dhlab.fbk.eu/Timeline_events/`.

2.1 First studies on events in the NLP community

In 2001, during the Workshop 'Temporal and Spatial Information Processing', three relevant works dealing with event annotation and processing were presented, each of them relying on a different notion of event. Filatova and Hovy (2001), whose system assigns a position on a timeline to events in newspaper articles, define events as propositions that contain a subject and a predicate. The system achieves a precision of 0.55 and a recall of 0.60. Schilder and Habel (2001) present a tool for the automatic annotation of temporal expressions and events in news. The authors define events as expressions that have an implicit time dimension and are either verbs or noun phrases. The list of markable nouns is limited to those directly connected to a temporal expression or a temporal preposition (e.g., *after the election in May*) and belonging to the domain of interest (i.e., finance, *opening of the stock exchange*). In a further extension of the system, the authors perform event recognition through an ontology containing event-denoting nouns in the financial domain and information on event types (Schilder and Habel 2003). Finally, Katz and Arosio (2001) propose a method to annotate temporal relations at sentence level, limiting events to verbs. The three works highlight the need to achieve a consensus on a definition of event, aimed also at making automatic approaches comparable.

In that same year, Setzer (2001) presents *Sheffield Temporal Annotation Guidelines*, the first annotation scheme that takes into account all temporal information elements (i.e., events, temporal expressions, temporal relations and event identity). The author defines an event as something that happens, must be anchorable in time, can be instantaneous or may last for a period of time. States are therefore not taken into consideration and, from the linguistic point of view, candidate events include nominalizations, finite and non-finite verbs. Each event is associated with attributes giving grammatical and semantic information, e.g., aspect.

Built upon Sheffield Temporal Annotation Guidelines, TimeML (Pustejovsky *et al.* 2003a) is a scheme for the annotation of events, temporal expressions and relations between events and/or temporal expressions (i.e. temporal, aspectual and subordination relations). Following Bach's broad notion of event, TimeML identifies a wide range of linguistic expressions realizing events, i.e. tensed and untensed verbs (e.g., *was captured*, *to thank*), adjectives (e.g., *sick*), nominals (e.g., *strike*) and prepositional phrases (e.g., *on board*). The consolidation of TimeML as an international standard called ISO-TimeML (ISO 2008) has facilitated its adaptation to different languages, such as Spanish (Saurí 2010) and Korean (Im *et al.* 2009) and the release of annotated data, such as the English TimeBank (Pustejovsky *et al.* 2003b) .

2.2 Evaluation campaigns

Parallel to the works reported in the previous Subsection, several evaluation campaigns on temporal IE and processing have been carried out. As shown in Figure 1, such campaigns have become very frequent in the last decade, with some years characterized by multiple evaluations.

The first campaign was the Message Understanding Conference (MUC-3) in 1991. It hosted the 'Scenario Template' (ST) task, in which systems were required to identify information about a given event (e.g., an air vehicle launch) and relate such information to the entities involved in it. Thus, an event was considered as a set of relationships between participants, time and space: From a practical point of view, it was seen as a template with slots to be automatically filled. The ST task was proposed in five MUC editions, from 1991 to 1998. Throughout the years, teams participating in ST presented systems with a modular pipeline architecture based mainly on pattern-matching techniques in particular after the success of the FASTUS system in MUC-4 that used such approach (Appelt *et al.* 1993). Results registered in the ST task are quite low if compared with the ones achieved in other MUC tasks such as Named Entity Recognition and Coreference (CO) resolution (Chinchor 1998). For example, in MUC-7, the best system in the ST task obtained 0.51 F-score (Aone *et al.* 1998), whilst the best systems in the Named Entity Recognition and CO tasks achieved an F-score of 0.93 and 0.62, respectively (Mikheev, Grover and Moens 1998; Hovy *et al.* 2013). The main difficulties of systems participating in the ST task were the complexity of texts to be processed, the high number of slots to be filled and the need of world knowledge for some of these slots.

In the 'Event Detection and Recognition' task, run for three years in the context of the ACE (Automatic Content Extraction) programme, an event is a specific occurrence involving participants, something that happens and can often be described as a change of state (Linguistic Data Consortium 2005). According to the ACE approach, extracting an event means marking up both the verb, noun, pronoun or adjective that most clearly expresses its occurrence (i.e., the event *trigger*) and the entire sentence containing that word (i.e., the event *mention*). However, only events belonging to a list of predefined types are taken into account, each with a number of subtypes (e.g., the event type Conflict has two subtypes: Attack and Demonstrate). Each event is associated with the entities playing a role in it (e.g., the location target of an Attack event) and a set of attributes such as genericity and tense.

It is not possible to make a precise comparison between ACE and MUC results because the former adopted a different evaluation measure called Value Score (Doddington *et al.* 2004). However, the two initiatives share the same limitation: They were both designed around specific domains and very limited types of events (Grishman 2010). Therefore, the proposed systems could hardly be adapted to different domains and applications. Another issue is that the corpora used for training and evaluation were artificially built by choosing the newspaper articles containing more events of interest: For example, forty-eight per cent of the events in the training corpus of ACE2005 belonged to the Attack subtype (Grishman 2010). This led to the creation of data sets that are not representative of journalistic language. Moreover, the complexity of ACE annotation makes the creation of consistent labeled data very challenging.

In order to address this last shortcoming, the ERE (Entities, Relations, Events) scheme has been developed within the DARPA DEFT programme (Aguilar *et al.* 2014), with the goal to propose a lighter weight version of ACE. ACE and ERE share the same definition of events and the same event ontology (thus, event

annotation is limited to the ACE types and subtypes). However, ERE simplifies the annotation by collapsing tags, accepting a looser event extent and reducing the set of attributes and values. Recently, a transition between this simple scheme (also known as Light ERE) towards a more sophisticated representation of events has been proposed under the name of Rich ERE (Song *et al.* 2015). In Rich ERE, the event ontology includes a new type and several new event subtypes. Moreover, the number of attributes is expanded and more attention is devoted to event CO. These DEFT ERE standards are the basis of the novel Event Nugget annotation scheme (Mitamura *et al.* 2015). An event nugget is a semantically meaningful unit referring to an event and linguistically represented not only by a single word but also by a continuous or discontinuous multi-token expression. The Knowledge Base Population evaluation track of the Text Analysis Conference (TAC KBP) conducted a task on event argument extraction and a pilot task on event nugget detection (Song *et al.* 2016) in 2014,[3] and these same tasks are included also in the Event Track of Knowledge Base Population evaluation track of the Text Analysis Conference 2015 and 2016.[4]

Although the Knowledge Base Population evaluation track of the Text Analysis Conference campaigns have been successful, their impact at large has been limited because the annotated datasets were distributed only to tasks participants. A different approach was adopted instead by TempEval organizers, who greatly contributed to improving state-of-the-art technologies in the field of Temporal Processing by making the data freely available after the campaigns. This consolidated also the success of TimeML annotation standard.

TempEval-1 (Verhagen *et al.* 2007) was the first open and international evaluation competition that used TimeBank as a benchmark. TempEval-1 avoids the complexity of complete temporal annotation by focussing only on the identification of temporal relations between given pairs of temporal expressions and events. TempEval-2 (Verhagen *et al.* 2010) was a more complex campaign than the previous one: It was multilingual and consisted of six subtasks including event extent identification and classification of event attributes. This subtask was proposed also in TempEval-3 (UzZaman *et al.* 2013). Only one out of seven participants in the event extraction and classification subtask uses a rule-based approach (Zavarella and Tanev 2013). The best performing systems rely on a supervised approach both for event extraction and event type classification: TIPSem (Llorens, Saquete and Navarro 2010), ATT1 (Jung and Stent 2013) and KUL (Kolomiyets and Moens 2013) are based on Conditional Random Fields, MaxEnt classification and Logistic Regression, respectively. They all take advantage of morphosyntactic information (e.g., POS) and semantic features at both the lexical and the sentence level, e.g., WordNet synsets (Fellbaum 1998) and semantic roles. Best results in event extraction are around 0.80 F1-score. However, when dealing with the classification of event types, system performances drop by almost ten points, with F1-scores all below 0.72.

[3] http://www.nist.gov/tac/2014/KBP/Event/index.html
[4] http://www.nist.gov/tac/2015/KBP/Event/index.html

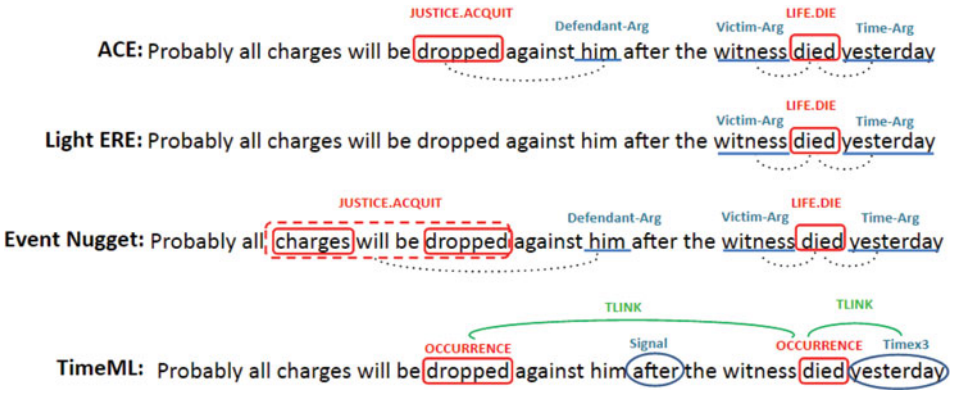

Fig. 2. (Colour online) Comparison of different event annotations. Red squares highlight event triggers whilst blue underlinings identify other annotated elements that in ACE, Light ERE and Event Nugget constitute event arguments. Connections between events and arguments are displays in dotted lines. For TimeML, temporal links are in green.

SemEval-2015 hosted three tasks related to temporal processing in the 'Time and Space track' with a focus on new challenges, new evaluation approaches and new domains.[5] The TimeLine task addressed CO resolution of events and temporal relation extraction at a cross document level with the aim of building timelines (Minard *et al.* 2015). QA TempEval introduced an extrinsic evaluation that took into consideration a specific end-user application, i.e., question answering (Llorens *et al.* 2015). Clinical TempEval moved past TempEval efforts from the news to the clinical domain (Bethard *et al.* 2015).

As a wrap-up of the different annotation schemes described in this section, we present in Figure 2 the same sentence annotated according to ACE, Light ERE, Event Nugget and TimeML guidelines. Differences in event types amongst ACE, Light ERE and Event Nugget are minimal (in this example are even null), whilst there is more variation concerning extent. ACE, Light ERE and TimeML annotate only events as single tokens, whilst Event Nugget annotation annotated multi-token and discontinuous expressions (*charges...dropped* in the third example). Moreover, in Light ERE, only actual events are eligible to be annotated (this is why *dropped* is not annotated in the second example). All the other schemes, instead, include the annotation of probable, possible and negated events. In ACE, Light ERE and Event Nugget events are connected to their arguments, i.e., entities such as *him* and *witness*. In TimeML, instead, the focus is on temporal links between two events (e.g., *dropped* and *died*) or between an event and a temporal expression (e.g., *died* and *yesterday*). In general, ACE, Light ERE and Event Nugget combine information on events with their argument structure, whilst in TimeML, the temporal dimension acquires more relevance, having its roots in Allen's interval algebra (Allen 1984).

[5] `http://alt.qcri.org/semeval2015/`

2.3 Adaptation of event processing to new domains

Most evaluation exercises presented so far were concerned with event processing in the news domain. Only recently, NLP researchers have started to look at different domains and develop domain-specific annotation guidelines and systems. For instance, following an increased interest in the temporal processing of clinical records, ISO-TimeML has been adapted to the clinical domain developing, as a result, the THYME annotation guidelines.[6] Following such guidelines, an event is 'anything relevant to the clinical timeline' (Styler *et al.* 2014), for example diseases, medical treatments and all actions and states related to the patient's clinical timeline. THYME guidelines formed the basis of both the i2b2 shared task in 2012 (Sun, Rumshisky and Uzuner 2013) and of the Clinical TempEval evaluation, organized within SemEval 2015 and aimed at assessing the performance of temporal IE systems on clinical notes and pathology reports.[7] The University of Colorado at Boulder proposed an extension of the THYME guidelines integrating ISO-TimeML, the Stanford Event CO (Lee *et al.* 2012) and the CMU Event CO guidelines (Hovy *et al.* 2013) under the name of Richer Event Description.[8] Richer Event Description adopts the TimeML wide definition of events and annotates events, temporal expressions and entities, as well as temporal, CO and casual relations (Hovy *et al.* 2013).

Three editions of the BioNLP shared task in 2009, 2011 and 2013 evaluated systems for extracting events from biomedical data. In this field, the definition of event is strongly domain-dependent and expert biologists annotate the datasets. More specifically, a biological event is a temporal occurrence involving one or more genes or proteins (Kim *et al.* 2006): An event ontology that defines a set of processes and functions supports the annotation. During the 2013 evaluation campaign, different tasks were proposed: In the Genia Event Extraction, task systems were required to detect trigger words expressing molecular and sub-cellular events (e.g., *mutation*), assign a type to each event (e.g., *anatomical* or *pathological*), link events to their arguments (e.g., a molecule) and identify speculated and negated events (e.g., the failure of a mutation) (Nédellec *et al.* 2013). EVEX, TEES-2.1, and BioSEM were the best performing systems in the extraction of events and of their primary arguments during BioNLP-ST 2013, with an F-score of 0.51. The first two systems combine Support Vector Machines and linguistic features, whilst the third one is rule-based (Hakala *et al.* 2013; Bjrne and Salakoski 2013; Bjrne *et al.* 2013).

Event extraction from social media is another emerging area of research (Atefeh *et al.* 2015). Most of the works in this field address the task as a clustering problem following the Topic Detection and Tracking approach mentioned in Section 1, for example, using an unsupervised method and focussing on the detection of unspecified new events (Petrović, Osborne and Lavrenko 2010). Other works deal with the retrieval of retrospective events in microblogs, such as Twitter: Amongst others, Metzler, Cai and Hovy 2012 propose a temporal query expansion technique to retrieve a ranked list of event summaries, having the events classified in different

[6] http://clear.colorado.edu/compsem/documents/
[7] http://alt.qcri.org/semeval2015/task6/
[8] https://github.com/timjogorman/RicherEventDescription

categories and types. Ritter, Mausam and Clark (2012) test a different approach appling IE techniques to identify events in a stream of tweet. The authors annotated manually event-referring phrases in a corpus of 1,000 tweets following the TimeML event definition and developed an automatic tagger that deals with the complexity of Twitter language (i.e., informal and ungrammatical style) achieving an F-score of 0.64.

The literature reports a number of works that try to tackle the semantics of historical texts using a combination of Semantic Web technologies and NLP approaches (Meroño-Peñuela *et al.* 2015). However, unlike what happened in the clinical domain, no attempt was made to find a domain-specific definition of event combining the historical perspective and ongoing research in the NLP field. Moreover, NLP techniques specifically developed for event processing have not been fully exploited and the current standardization efforts have received little attention in this domain. For example, in the *Agora project*, Van Den Akker *et al.* (2010) aimed at enriching museums metadata through the extraction of historical event names from unstructured texts, event extraction is assimilated to the recognition of named entities. Therefore, only named events, such as *French Revolution*, are taken into account. Another choice usually made in projects dealing with historical documents is limiting the extraction of events to a set of specialized types. For example, in the *FDR/Pearl Harbor project* (Ide and Woolner 2004) research focussed only on *communication* events. This choice was driven by the goal of the project, which was to help historians of WWII to search and retrieve information from documents (e.g., government correspondence and memoranda) written before the Pearl Harbor attack in 1941. This specific set of events was categorized based on FrameNet (Baker, Fillmore and Lowe 1998) by assigning verbs in the corpus of reference to the 'Communication' frame and its sub-frames. Another project, the *Semantics of History*, focusses only on *conflict-related* and *motion* actions (Cybulska and Vossen 2011).

Another weak point of current NLP research for historical texts is the scarcity of corpora fully annotated with temporal information. For example, files tagged within the projects described above have not been publicly released. Two notable exceptions are the ModeS TimeBank (Nieto, Saurí and Bernabé 2011), containing Spanish texts from the eighteenth century, and the De Gasperi corpus, a collection of documents written by the Italian statesman Alcide de Gasperi and dating back to the beginning of the twentieth century. Both were manually annotated following a language-specific adaptation of TimeML. ModeS TimeBank was employed for theoretical studies on the evolution of the Spanish language, whilst the De Gasperi corpus was used to measure the performance of event extraction systems on historical texts within the EVENTI evaluation exercise.[9]

In order to measure how an event extraction system trained on news performs on historical data, we evaluated the performance of two supervised systems (one for Italian and one for Spanish) trained on news and tested on in-domain and out-of-domain data, i.e. historical texts. For Italian, we used the state-of-the-art

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

FBK-HLT-Time system (Mirza and Minard 2014) evaluated both on contemporary news and on the De Gasperi corpus in the framework of the EVENTI evaluation campaign for event extraction. The system achieves **F1 0.87** (P 0.88, R 0.85) on news and **F1 0.83** (P 0.89, R. 0.78) on historical documents, showing that the different performance in the two domains is limited to a drop of 0.07 points in recall. This is mainly due to the fact that the language in the De Gasperi corpus is very similar to contemporary Italian, and tokens corresponding to events are generally easy to recognize. The evaluation on Spanish data, instead, is remarkably different. We evaluated the TIPSem system (Llorens *et al.* 2010) on the Modes TimeBank, and we compared it with the TempEval 2013 results of the same system (UzZaman *et al.* 2013). In this case, the performance in the two domains is very different: On news data, the system achieves **F1 0.89** (P 0.92, R 0.86), whilst its performance drops to **F1 0.39** (P 0.27, R 0.72) on the ModeS TimeBank, since the corpus shows many diachronical language variations affecting precision.

This comparison shows that event recognition systems trained on news in some cases would be suitable for investigations in new domains, given that the event definition framework and the language of the documents to be processed are similar to the ones used for training. For instance, historians analysing corpora dating back to the previous century may still achieve satisfactory system performance.

In order to account for all corpora annotated so far with event information in different domains and languages, we report a summary in Table 1. The information presented in the table was gathered through the direct analysis of the resources downloaded from the Web and merging data from scientific papers. Resources listed in the table have been annotated following different schemes and cover five domains, with a prevalence of the news domain. The number of corpora in the list shows the interest of the NLP community in event processing. The most recent corpora confirm the trend towards new domains, new languages and more complex tasks integrating event extraction.

3 What is an event in history?

As shown in the previous Section, past projects trying to apply NLP techniques to historical investigation have adopted heterogeneous approaches, and there has been no real effort amongst history scholars to standardize event definition taking into account past proposals made in the NLP community. However, researchers in history face daily issues related to the observation, analysis and interpretation of events. This gap between the two research communities may depend on a lack of communication and cross-fertilization, but also on the fact that events as defined in IE do not fully satisfy requirements from other disciplines. In order to clarify the reasons of this gap, we ran an investigation involving historians based on an online questionnaire. We circulated the survey in English and Italian, to facilitate the inclusion of different communities in this study. Both versions had the same set of questions. Only the examples taken from historical documents were different but with the same range of linguistic phenomena to investigate. The questionnaire was distributed via social media (i.e., Twitter and LinkedIn),

Table 1. *Corpora including event annotation in different domains. For each corpus the language, number of tokens, number of files and number of annotated events are provided. The symbol '-' is used in case of missing information. Resources in boldface are available online at the moment of writing*

Domain	Corpus	Lang	#Tokens	#Files	#Events
News	**ACE 2005 (training)**[*]	EN	259,889	599	4167
		ZH	307,991	633	3332
	French TimeBank (Bittar *et al.* 2011)	FR	15,423	109	2,115
	Romanian TimeBank (Forascu and Tufi 2012)	RO	65,375	181	7,926
	TimeBankPT (Forascu and Tufi 2012)	PT	69,702	182	7,887
	Persian TimeBank (Yaghoobzadeh *et al.* 2012)	FA	26,949	43	4,237
	Catalan TimeBank 1.0[†]	CA	75,800	210	12,342
	Spanish TimeBank 1.0[‡]	ES	75,800	210	12,641
	BCCWJ-TimeBank (Asahara *et al.* 2013)	JA	56,518	54	3,824
	EVENTI corpus (Caselli *et al.* 2014)	IT	130,279	366	21,633
	TempEval 1 (training)[§]	EN	52,740	162	5,150
	TempEval 2 (training+test)[‖]	ZH	32,788	61	1,204
		EN	62,613	184	2,256
		IT	31,995	66	1,036
		FR	13,387	98	248
		KO	16,900	28	602
		ES	56,880	212	2,129
	TempEval-3 (AQUAINT+TimeBank+Platinum)[¶]	EN	102,375	276	12,534
	FactBank (Saurí and Pustejovsky 2009)	EN	77,000	208	9,500
	2012 EventCorefBank (ECB) (Lee *et al.* 2012)	EN	–	482	2,533
	ECB+ (Cybulska and Vossen 2014)	EN	377,367	982	15,003
	Light ERE[#] (Mott *et al.* 2016)	ZH[**]	127,458	171	481
		EN	101,191	171	369
	Rich ERE[#] (Mott *et al.* 2016)	ZH[**]	127,458	171	1,491
		ES	101,191	171	2,933
	Event Nugget (training+test) (Mitamura *et al.* 2015)	EN	336,126	351	10,719
	TimeLine (Minard *et al.* 2015)	EN	29,893	90	915
	MEANTIME[††] (Minard *et al.* 2016)	EN	13,981	120	2,096
		IT	15,676	120	2,208
		ES	15,843	120	2,223
		NL	14,647	120	2,223
Clinical	i2b2 (Sun *et al.* 2013)	EN	178,000	349	30,000
	Clinical TempEval (Train+Dev) (Bethard *et al.* 2015)	EN	533,393	440	59,864
Biomedical	**GENIA** (Kim, Ohta and Tsujii 2008)	EN	–	1,000	36,114
Social media	Twitter NLP (Ritter *et al.* 2012)	EN	19,484	1,000	–
History	**ModeS TimeBank**[‡‡]	ES	25,611	102	1,261
	De Gasperi Corpus (Caselli *et al.* 2014)	IT	5,671	10	1,195

[*] `https://catalog.ldc.upenn.edu/LDC2006T06`
[†] `https://catalog.ldc.upenn.edu/LDC2012T10`
[‡] `https://catalog.ldc.upenn.edu/LDC2012T12`
[§] `http://www.timeml.org/tempeval/`
[‖] `http://timeml.org/tempeval2/`
[¶] `http://www.cs.york.ac.uk/semeval-2013/task1`
[#] Light ERE, Rich ERE and Event Nugget corpora include both news and discussion forum data
[**] Number of characters instead of the number of tokens
[††] `http://www.newsreader-project.eu/results/data/wikinews/`
[‡‡] `https://catalog.ldc.upenn.edu/LDC2012T01`

Table 2. *English sentences annotated by the questionnaire participants. For each sentence, we report the absolute percentage of annotated events in terms of single tokens (ST), multi-token expressions (MT), verbal expressions (V) and non verbal expressions (NV). The three most common extents for each sentence are also reported. For Italian, we registered comparable results*

Sentences		ST	MT	Most common extents
Today, once again, the independence of the Western Hemisphere is menaced from abroad	**V**	17%	8%	today
	NV	51%	24%	menaced independence
This country has not been prepared for any disarmament, arms control or atomic testing conference that has taken place since the end of the Korean war	**V**	4%	13%	conference end of the
	NV	28%	55%	Korean war disarmament
I think we can work that out with the advice of the Ways and Means Committee	**V**	20%	28%	advice
	NV	36%	16%	work that out think

mailing-lists (e.g., the Humanist Discussion Group) and targeted emails to individual historians, professional associations (e.g., the Australian Historical Association) and research centres (e.g., Institute of Historical Research at the University of London). After two months from its launch, seventy-four historians participated in the survey.[10] The expertise of the survey respondents covered many research fields such as cultural history, history of science, political science and biography thus providing insight into current research practices in the domain on interest.

The general goal of this analysis, that can likely be applied also to other domains, was to leverage knowledge about the way events are defined in historical research and to compare it with ongoing standardization efforts in the NLP community. To the best of our knowledge, this is the first questionnaire on this topic, whose outcome can potentially enrich the current theoretical discussion on the nature of events. Besides, it can be seen as a preliminary step towards the definition of annotation guidelines for developing NLP tools in this domain.

3.1 Questionnaire description and results

Two parts composed the questionnaire. In the first one, we collected participants' demographic information and assessed their general interest in NLP. The second part aimed at shedding light on the notion of 'event' for historians based on three questions.

In the first question, participants were asked to list all the single words or expressions encoding events (if any) in three given sentences, without providing any definition of what an event is. The aim was to indirectly leverage an operational definition of events based on historians' knowledge.

[10] Half of the participants filled in the questionnaire in Italian and half in English.

The sentences were different in the English and in the Italian questionnaire but they contained the same linguistic phenomena: negated verbs (e.g., *has not been prepared*), nominalizations (e. *disarmament*), aspectual nominals (e.g., *end*), cognitive verbs (e.g., *think*), named events (e.g., *Korean war*), nominals expressing states (e.g., *independence*) and multi-token expressions like phrasal verbs (e.g., *taken place*). We report in Table 2 the English sentences, taken from J. F. Kennedy's public speeches.[11] Questionnaire participants could annotate single words or expressions conveying events, but also provide no annotation. In the first and the third sentence, a high percentage of respondents (thirty-seven per cent and sixty-eight per cent, respectively) did not detect any event, probably because these sentences contain a state (i.e.,*independence*) and an opinion (i.e., *I think…*). The second sentence includes a named event (*Korean war*) and only the three per cent of respondents did not annotate any event.

The percentages listed in Table 2 are calculated by taking into consideration the total number of annotations per sentence. To decide if an expression was verbal or non-verbal, we looked at the part of speech of the words contained in the expression. In particular, all expressions containing at least one verb were considered verbal (e.g., *menaced*, *work that out*, *think*). Anyway the majority of the identified events are non-verbal (e.g., *today*, *independence*, *conference*, *end of the Korean war*, *advice*): seventy-five per cent in the first sentence, eighty-three per cent in the second and fifty-two per cent in the third. This contrasts with the outcome of the experiment reported by Hatzivassiloglou and Filatova (2003), in which nouns such as *war* and *earthquake* were never identified as events by a group of students annotating news. Events consisting of more than one token are annotated very frequently in all the sentences: thirty-two per cent of events annotated in the first sentence are multi-token, sixty-eight per cent in the second and forty-four per cent in the third. Some of these multi-token extents correspond to entire clauses, e.g., *This country has not been prepared* and *I think we can work that out*. This high number of multi-token events goes against the TimeML and Richer Event Description minimal chunk rule for tag extent, according to which only single tokens are to be annotated as events.[12] The distinction made in ACE and ERE between event trigger (the word expressing the event) and event mention (the sentence containing it) seems to better meet historians' needs. Moreover, ACE, ERE and Event Nugget allow the annotation of multi-token event triggers (the latter also discontinuous cases).

Conclusion 1. The notion of event is seen as independent from its grammatical category, in line with TimeML. However, the minimal chunk annotation used in TimeML is not optimal. Amongst the considered standards, the multi-token annotation of continuous and

[11] Available at `http://www.presidency.ucsb.edu/1960_election.php`.

[12] The only exception to the minimal chunk rule present in the TimeML guidelines is given by exocentric predicative elements for which the entire expression is to be annotated (*All seventy-five people were **on board** at 9:00 a.m.*). ISO-TimeML contains a very generic sentence that leave space for other exceptions thus it seems that the need for multi-token events is taken into consideration by researcher working on the TimeML definition. However, there is no evidence that a concrete step has been made in this direction for English. On the contrary, in some adaptations of TimeML to other languages (e.g., It-TimeML), multi-token annotation is allowed (Caselli *et al.* 2014).

discontinuous multi-token expressions proposed in Event Nugget addresses best historians' view on events.

In the second question, we asked participants to rate the relevance of a list of properties to define when a word or expression can be labelled as an event. These properties included for instance impact, cause and frequency, and were inspired by the essay 'What is an Event?' written by the history scholar Robert Bedrosian.[13] The ratings included four possible values, i.e., 'very important', 'somewhat important', 'not important' and 'don't know'. Figure 3 presents the value distribution across the properties merging results from English and Italian questionnaires. *Public Perception* and *Impact*, i.e., the degree to which an event affects society or nature, are properties not related to the linguistic analysis of texts but to the historians' interpretative work. Both were considered quite relevant, especially the latter. *Predictability* is the only property in which the value 'not important' prevails. On the contrary, *Type* has the highest positive consensus. In TimeML, event type information is conveyed by seven possible values of the `class` attribute, where both semantic (e.g., `STATE`) and syntactic criteria (e.g., `I_STATE`) are taken into account. A classification based on syntactic criteria would not be optimal for historians, for whom syntax does not have a primary importance in the interpretation of the text. On the other hand, the event ontology of ACE, ERE and Event Nugget is made of a list of types and subtypes which limits the annotation to a specific set of categories strongly connected to the news domain (e.g., type: `JUSTICE`, subtype `ACQUIT` in Figure 2). Other categories should be added to this ontology to make it more apt for the history domain so to include, for example, events of cognition, emotion and communication. The USAS (Rayson *et al.* 2004) and the Historical Thesaurus of English tagsets (Kay *et al.* 2009) contain twenty-one and thirty-seven main semantic categories, respectively and they have been already used to analyse historical texts (Archer 2014; Rayson *et al.* 2015): For this reason, they can provide an interesting fine-grain classification of events for the history domain. *Factuality*, i.e., the distinction between actual real facts and imaginary, future, avoided and prevented events, has a limited interest for historians, whilst it is more relevant from a linguistic perspective. In fact, TimeML encodes this property through subordination links whereas other annotation schemes encode it as an attribute attached to the event (Saurí and Pustejovsky 2009; van Son *et al.* 2014). *Preceding and consequent events* appear to be very important for historians, and this is in line with the ongoing effort in NLP to encode intra- and cross-document event ordering. TimeML conveys this information by using temporal links, corresponding to thirteen types of binary temporal relations, inherited from Allen's interval algebra. Besides, the challenge of cross-document event ordering has been recently addressed by the TimeLine task at SemEval-2015.[14] In TimeML, the temporal link tag is also employed to link events to points in time (e.g., *25/12/2014*), durations (e.g., *3 month*) and temporal expressions denoting recurring times (e.g., *every month*): This corresponds to the *Temporal Grounding* property, that is the

[13] `http://rbedrosian.com/event.htm`
[14] `http://alt.qcri.org/semeval2015/task4/`

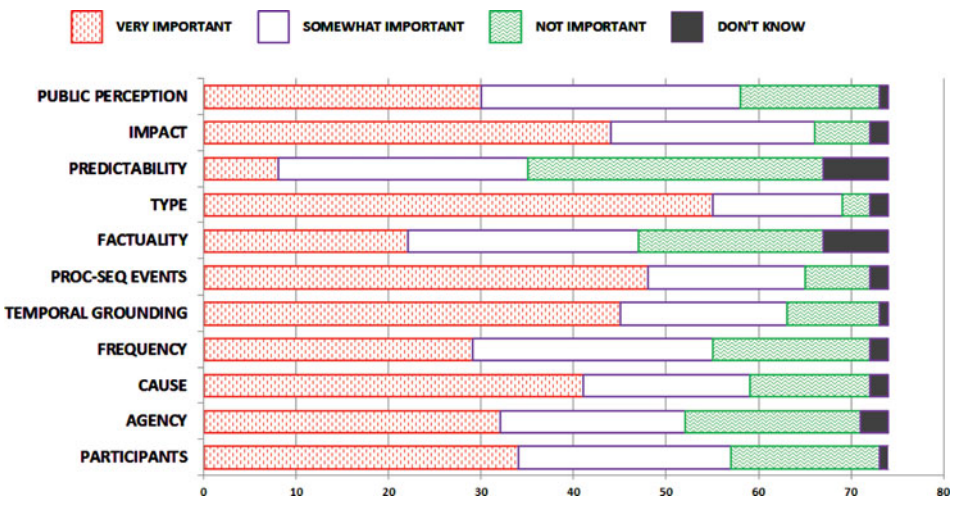

Fig. 3. (Colour online) What are the most important properties for a historian in order to understand if a word (or a set of words) expresses a relevant event.

degree to which an event can be pinpointed to a particular time or period, and the *Frequency* property. In the MUC ST as well as in ACE, ERE and Event Nugget, temporal relations between events or between an event and a temporal expression are not explicitly addressed. The link between an event and a temporal expression is encoded in the form of a temporal slot in case of MUC or of a temporal argument in case of ACE, ERE and Event Nugget (e.g., the *Time-Arg* argument 'yesterday' of the event trigger 'died' in Figure 2). The property of an event being the cause or the effect of another event (i.e., *Cause*) is strictly connected to the *Agency* property, i.e., who/what caused such event. TimeML does not include a specific relation for causative constructions but causes and effects denoted by events are temporally ordered using a temporal link (a cause always precedes the effect). However, attempts have been made to explicitly annotate causal relations as an extension of TimeML (Mirza and Tonelli 2014). In ACE, ERE and Event Nugget, *Agency* is annotated as event argument for several event types. For example, in the sentence 'his father-in-law killed him', *father-in-law* is the Agent argument of the trigger event *killed* of type LIFE. Event–event causality relations are planned as future development of the Rich ERE annotation, but they are currently not included in the guidelines. On the contrary, causal relations play an important role in the Richer Event Description guidelines (Hovy *et al.* 2013). As for *Participants*, TimeML does not foresee the annotation of the entities involved in an event, even if historians' responses suggest that this information is quite relevant. Attempts have been made to add participants' information to events (Pustejovsky, Littman and Saurí 2007), but this has not led to the extension of TimeML specifications. On the contrary, participants annotation is crucial in MUC, ACE, ERE and Event Nugget, in which several arguments have to be identified (e.g., *Victim-Arg* in Figure 2). Research on semantic roles can provide much guidance in this respect, for example, by taking inspiration from PropBank (Palmer, Gildea and Kingsbury 2005) or FrameNet frameworks. This was already proposed within the NewsReader project (Vossen *et al.* 2014), where event extraction

from news is performed by leveraging information related to events and participants from different sources and modelling them as knowledge graphs. An analysis of the application of NewsReader pipeline to the history domain may provide useful insights into the importance of semantic roles from historians' perspective.

Conclusion 2. An event is a complex information object characterized by many properties. A new framework for the annotation of events in historical texts should take advantage of the temporal dimension as defined in TimeML but also look at other annotation efforts (e.g., semantic roles in FrameNet, participants' information in Event Nugget) to cover all important properties of events.

Finally, in the third question, participants were asked to choose between two linguistic annotations of short text snippets containing the same specific phenomena in both English and Italian, i.e., states and multi-token expressions. This question had the aim of confirming or disproving the previous conclusions. The English questionnaire presented the following passage taken from a speech uttered by J.F. Kennedy:

After the key African state of Guinea, now $[voting]_{1-2}$ *with the Soviet Union in Communist foreign policy, after it* $[gained]_{1-2}$ *its* $[independence]_2$, *a Russian Ambassador* $[[showed]_1up]_2$ *the next day. Our Ambassador did not* $[[show]_1up]_2$ *for 9 months.*

In the annotation marked with $[\ldots]_1$, only single tokens are annotated as events following the TimeML specifications. Moreover the state *independence* is not annotated. The option marked with $[\ldots]_2$ proposes looser criteria, annotating both multi-token event expressions and states. Only five per cent of participants preferred the first annotation, sixty-one per cent chose the second option and the rest did not give preference to either of the two annotations. We asked for the motivations behind this choice: Respondents said that a broad context is needed to represent events (An event is not one word, it's syntactical, inter-relation between agent and object/patient). Besides, answers highlighted the importance of states and conditions (I feel that the state/condition is important.). In ACE, ERE and Event Nugget, states that result from actions, such as being dead, married or retired, are included in the annotation, but disagreement is an open issue for human annotators (Mitamura *et al.*, 2015). On the other hand, in TimeML, only states that are temporally relevant (e.g., that are bound to a specific point or period of time) have to be annotated. Defining what states have to be annotated using a predefined set of annotation rules, as in TimeML, would be extremely critical because such rules could not cover all the information needs of historians.

Conclusion 3. Conclusion 1 about multi-token annotation is confirmed, showing that TimeML could not be applied to a new domain as is. Moreover, states/conditions are important and should be considered in the annotation of historical documents.

4 Conclusions

This paper presents a survey of the state of the art in event definition and processing in NLP, adopting an inter-disciplinary perspective. In the last twenty-five years, thanks to many workshops and evaluation exercises dedicated to the semantic

and linguistic analysis of events, research has moved forward. However, a careful adaptation of existing annotation schemes is necessary to apply the outcome of these research activities to new domains. On the basis of the analysis of the state of the art and of historians' replies to our questionnaire, we can now answer the questions posed in Section 1:

(i) *Was all the work devoted to event processing with IE techniques useful to serve real historical investigation?* NLP methods and technologies have not been fully exploited yet in the domain of history. Existing annotation schemes and systems constitute an important starting point but a careful adaptation is necessary to meet the requirements of domain experts. In the near future, we plan to propose an annotation paradigm for events that takes into account both historians' suggestions collected through the survey and past experience in NLP.

(ii) *Were the various definitions of events provided over the years compatible with research practices adopted in other communities?* Several event definitions have been proposed over the years, each showing specific strengths and weaknesses. TimeML event definition relies on the broad notion of eventuality: The fact that it includes states as well as processes and actions is compatible with historians' needs. On the other hand, states should be taken into consideration even if not bound to a specific point or period of time. Allowing only single token events does not meet research practices adopted in other domains. The multi-token choice proposed in the Event Nugget initiative addresses better this need.

(iii) *How should events be defined to be processable with NLP tools but also to comply with historical research?* Events can be defined as complex information objects characterized by many properties. These can be cast by combining different NLP analyses providing rich semantic information, such as semantic role labelling, causality detection and temporal relation processing. The role played by this information in historical research, however, can vary a lot according to the historiographical approach used. For example, the so-called *evenemential* approach defines history as a chronological accumulation of events in a coherent timeline (Simiand 1960). From this perspective, events are objective entities, atomic facts that do not need deep interpretation. In sharp contrast with this approach, more recent theories propose looking at events in a long-term perspective (Guldi and Armitage 2014), in order to study them in their connection with other events taking into consideration recurring analogies and structures. Following these last assumptions, historians have the duty to pose problems and formulate hypotheses, not only to observe events emerged from the analysis of historical documents but also to interpret them (Febvre 1953). The distinction between an important event and one with no historical value is thus never definitive because the research question changes constantly according to the documents that historians are analysing (Marrou 1954). The challenge is to develop an annotation scheme and subsequently a system

to support historical investigation and guide the interpretation of historians without replacing them or influencing their understanding of documents.

References

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and Ellis, J. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Baltimore, Maryland, USA: ACL, pp. 45–53.

Allen, J. F. 1984. Towards a general theory of action and time. *Artificial Intelligence* **23**(2): 123–54.

Aone, C., Halverson, L., Hampton, T., and Ramos-Santacruz, M. 1998. SRA: description of the IE2 system used for MUC-7. In *Proceedings of the 7th Message Understanding Conferences (MUC-7)*. Fairfax, VA.

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., and Tyson, M. 1993. FASTUS: a finite-state processor for information extraction from real-world text. In *Proceedings of IJCAI*, Chambéry, France, vol. 93, pp. 1172–1178.

Archer, D. 2014. Exploring verbal aggression in English historical texts using USAS. In I. Taavitsainen, A. H. Jucker and J. Tuominen (eds.), *Diachronic Corpus Pragmatics (Pragmatics & Beyond New Series)*, pp. 273–302. University of Helsinki / University of Zurich: John Benjamins Publishing Company.

Asahara, M., Yasuda, S., Konishi, H., Imada, M., and Maekawa, K. 2013. BCCWJ-TimeBank: temporal and event information Annotation on Japanese text. In *Proceedings of PACLIC 27*, Taipei, Taiwan, pp. 206–214.

Ashish, N., Appelt, D., Freitag, D., and Zelenko, D. 2006. Papers from the AAAI workshop on event extraction and synthesis. Technical Report WS-06-07, American Association for Artificial Intelligence.

Bach, E. 1986. The algebra of events. *Linguistics and Philosophy*, **9**(1): 5–16, D. Reidel Publishing Company.

Baker, C. F., Fillmore, C. F., and Lowe, J. B. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montréal, Quebec, Canada: ACL, pp. 86–90.

Berr, H. 1922. L'histoire traditionnelle et la synthse historique. *Revue belge de philologie et d'histoire* **1**(3): 556–556.

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., and Verhagen, M. 2015. SemEval-2015 Task 6: clinical TempEval. In *Proceedings of SemEval 2015*, Denver, Colorado, USA: ACL.

Bittar, A., Amsili, P., Denis, P., and Danlos, L. 2011. French TimeBank: an ISO-TimeML annotated reference corpus. In *Proceedings of ACL*, Portland, Oregon, USA, ACL, pp. 130–134.

Bjrne, J., and Salakoski, T. 2013. TEES 2.1: automated annotation scheme learning in the BioNLP 2013 Shared Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria.

Bui, Q. C., Campos, D., van Mulligen, E. M., and Kors, J. A. 2013. A fast rule-based approach for biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria.

Caselli, T., Bartalesi Lenzi, V., Sprugnoli, R., Pianta, E., and Prodanof, I. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, Portland, Oregon: ACL.

Caselli, T., Sprugnoli, R., Speranza, M., and Monachini, M. 2014. EVENTI EValuation of events and temporal INformation at Evalita 2014. In *Proceedings of the 4th International Workshop EVALITA 2014*, Pisa, Italy, Pisa University Press.

Chinchor, N. A. 1998. Overview of MUC-7/MET-2. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Fairfax, VA.

Cybulska, A., and Vossen, P. 2011. Historical event extraction from text. In *Proceedings of the 5th ACL-HLT LaTeCH Workshop*, Portland, Oregon: ACL, pp. 39–43.

Cybulska, A., and Vossen, P. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of LREC 2014, pp. 26–31*, Reykjavik, Iceland, European Language Resources Association (ELRA), pp. 26–31.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of LREC 2004*, Lisbon, Portugal, European Language Resources Association (ELRA).

Farzindar, A., and Khreich, W. 2015. A survey of techniques for event detection in Twitter. *Computational Intelligence* **31.1**(2015): 132–164.

Febvre, L. P. V. 1953. *Combats pour l'histoire*. Paris: Armand Colin.

Fellbaum, C. 1998. WordNet. *Blackwell Publishing Ltd*. Cambridge: MIT Press.

Filatova, E., and Hovy, E. 2001. Assigning time-stamps to event-clauses. In *Proceedings of the ACL-EACL 2001 Workshop for Temporal and Spatial Information Processing*, Toulouse, France: ACL.

Forascu, C., and Tufi, D. 2012. Romanian TimeBank: an annotated parallel corpus for temporal information. In *Proceedings of LREC 2012*, Istanbul, Turkey. European Language Resources Association (ELRA).

Grishman, R. 2010. The impact of task and corpus on event extraction systems. In *Proceedings of LREC 2010*, Valletta, Malta, European Language Resources Association (ELRA).

Guldi, J., and Armitage, D. 2014. *The History Manifesto*. Cambridge: Cambridge University Press.

Hagège, C. 1996. *L'homme de Paroles: Contribution Linguistique aux Sciences Humaines*. Fayard, Paris.

Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., and Ginter, P. 2013. EVEX in ST13: application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria.

Hatzivassiloglou, V., and Filatova, E. 2003. Domain-independent detection, extraction, and labeling of atomic events. In *Proceedings of RANLP*, Borovetz, Bulgaria, pp. 145–152.

Hovy, E., Mitamura, T., Verdejo, F., Araki, J., and Philpot, A. 2013. Events are not simple: identity, non-identity, and quasi-identity. In *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Atlanta, Georgia, USA: ACL.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. 1998. University of Sheffield: description of the LaSIE-II system as used for MUC-7. In *Proceedings of the 7th Message Understanding Conferences (MUC-7)*. Fairfax, VA.

Ide, N., and Woolner, D. 2004. Exploiting semantic web technologies for intelligent access to historical documents. In *Proceedings of LREC 2004*, Reykjavik, Iceland, European Language Resources Association (ELRA).

Ikuta, R., Styler IV, W. F., Hamang, M., O'Gorman, T., and Palmer, M. 2014. Challenges of adding causation to Richer Event Descriptions. In *Proceedings of the The 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Baltimore, Maryland, USA: ACL.

Im, S., You, H., Jang, H., Nam, S., and Shin, H. 2009. Ktimeml: specification of temporal and event expressions in korean text. In *Proceedings of the 7th Workshop on Asian Language Resources*, Suntec, Singapore: ACL, pp. 115–122.

Jung, H., and Stent, A. 2013. Att1: temporal annotation using big windows and rich syntactic and semantic features. In *Proceedings of * SEM*, Atlanta, Georgia, USA, vol. 2, pp. 20–24.

Kay, C., Roberts, J., Samuels, M., and Wotherspoon, I. 2009. Unlocking the OED: the story of the historical thesaurus of the OED. In *Historical Thesaurus of the Oxford English Dictionary: With Additional Material from a Thesaurus of Old English*. Oxford: Oxford University Press.

Katz, G., and Arosio, F. 2001. The annotation of temporal information in natural language sentences. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*, Toulouse, France: ACL.

Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. 2006. GENIA corpus manual. Technical Report, Citeseer.

Kim, J. D., Ohta, T., and Tsujii, J. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* **9**(1): 10.

Kolomiyets, O., and Moens, M. F. 2013. KUL: a data-driven approach to temporal parsing of documents. In *Proceedings of SemEval 2013*, Atlanta, Georgia, USA, pp. 83–87.

Lee, H., Recasens, M., Chang, A., Surdeanu, M., and Jurafsky, D. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of EMNLP-2012*, Jeju, South Korea, pp. 489–500.

Linguistic Data Consortium. 2005. ACE (automatic content extraction) english annotation guidelines for events, version 5.4.3 2005.07.01.

Llorens, H., Chambers, N., UzZaman, N., Mostafazadeh, N., Allen, J., and Pustejovsky, J. 2015. SemEval-2015 task 5: QA TEMPEVAL-Evaluating temporal information understanding with question answering. In *Proceedings of SemEval 2015*, Denver, Colorado: ACL.

Llorens, H., Saquete, E., and Navarro, B. 2010. TIPSem (English and Spanish): evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of SemEval-2010*, Uppsala, Sweden: ACL, pp. 284–291.

Marrou, H.-I. 1954. De la connaissance historique. Paris: Seuil.

Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., and van Harmelen, F. 2015. Semantic technologies for historical research: a survey. *Semantic Web Journal*, **6**(6): 539–64. IOS Press.

Metzler, D., Cai, C., and Hovy, E. 2012. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada: ACL.

Mikheev, A., Grover, C., and Moens, M. 1998. Description of the LTG system used for MUC-7. In *Proceedings of 7th Message Understanding Conference (MUC-7)*, Fairfax, VA.

Minard, A. L., Speranza, M., Agirre, E., Aldabe, I., van Erp, M., Magnini, B., Rigau, G., and Urizar, R. 2015. Semeval-2015 task 4: timeline: cross-document event ordering. In *Proceedings of SemEval 2015*, ACL. Denver, Colorado, USA: Association for Computational Linguistics.

Minard, A. L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of LREC 2016*, Portoro, Slovenia, European Language Resources Association (ELRA).

Mirza, P., and Minard, A. L. 2014. FBK-HLT-time: a complete Italian temporal processing system for EVENTI-Evalita 2014. In *Proceedings of the 4th International Workshop EVALITA 2014*, Pisa, Italy: Pisa University Press.

Mirza, P., and Tonelli, S. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014*, Dublin, Ireland: Dublin City University and ACL, pp. 2097–2106.

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., and Strassel, St. 2015. Event nugget annotation: processes and issues. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Denver, Colorado, USA: ACL, pp. 66–76.

Mott, J., Bies, A., Song, Z., and Strassel, S. 2016. Parallel Chinese-English entities, relations and events corpora. In *Proceedings of the 10th International Conference on Language*

Resources and Evaluation (LREC 2016), Portoro, Slovenia, European Language Resources Association (ELRA).

Nédellec, C., Bossy, R., Kim, J. D., Kim, J. J., Ohta, T., Pyysalo, S., and Zweigenbaum, P. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, pp. 1–7.

Nieto, M. G., Saurí, R., and Bernabé Poveda, M. A. 2011. ModeS TimeBank: a modern spanish TimeBank corpus. *Procesamiento del lenguaje natural* **47**(2011): 259–267.

Palmer, M., Gildea, D., and Kingsbury, P. 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*, **31**(1): 71–106.

Petrović, S., Osborne, M., and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *Proceedings of NAACL 2010*, Los Angeles, California, USA, pp. 181–189.

Pustejovsky, J. 1991. The syntax of event structure. *Cognition* **41**(1–3): 47–81.

Pustejovsky, J. 2005. A survey of dot objects. Author's weblog. Retrieved from URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.208.7525&rep=rep1&type=pdf`

Pustejovsky, J., Castaño, J. M., Ingria, R., Saurí, R. G., Setzer, A. and Katz, G. 2003. TimeML: robust specification of event and temporal expressions in text. In *Proceedings of IWCS-5*. Tilburg, The Netherlands.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. 2003. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, Lancaster, UK, pp. 647–656.

Pustejovsky, J., Littman, J., and Saurì, R. 2007. Arguments in TimeML: events and entities. In Schilder, F., Katz, G., & Pustejovsky, J. (eds), *Annotating, Extracting and Reasoning about Time and Events: International Seminar Revised Papers (Lecture Notes in Computer Science)*, pp. 107–126. Berlin: Springer Berlin Heidelberg.

Rayson, P., Archer, D., Piao, S., and McEnery, A. M. 2004. The UCREL semantic analysis system. In *Proceedings of the Beyond Named Entity Recognition Semantic Labelling for NLP Tasks Workshop*, Lisbon, Portugal, pp. 7–12.

Rayson, P., Baron, A., Piao, S., and Wattam, S. 2015. Large-scale time-sensitive semantic analysis of historical corpora. In *Proceedings of the 36th Meeting of ICAME*, Trier, Germany.

Ritter, A., Mausam, E. O., and Clark, S. 2012. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China: ACM, pp. 1104–1112.

Sasse, H. J. 2002. Recent activity in the theory of aspect: accomplishments, achievements, or just non-progressive state. *Linguistic Typology* **6**(2): 199–271.

Saurı, R. 2010. Annotating temporal relations in catalan and spanish TimeML annotation guidelines. Technical Report BM 2010-04, Barcelona Media.

Saurí, R., and Pustejovsky, J. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation* **43**(3): 227–68.

Schilder, F., and Habel, C. 2001. From temporal expressions to temporal information: semantic tagging of news messages. In *Proceedings of the ACL-EACL 2001 Workshop for Temporal and Spatial Information Processing*. Toulouse, France: ACL.

Schilder, F., and Habel, C. 2003. Temporal information extraction for temporal question answering. In *New Directions in Question Answering*, AAAI, pp. 35–44.

SemAf/Time Working Group ISO, 2008. *ISO DIS 24617-1: 2008 Language Resource Management - Semantic Annotation Framework - Part 1: Time and Events*. ISO Central Secretariat, Geneva.

Setzer, A. 2001. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. PhD Thesis, University of Sheffield.

Simiand, F. 1960. Méthode historique et science sociale. In Annales. *Histoire, Sciences Sociales* **15**(1): 83–119. France: EHESS.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. 2015. From light to rich ERE: annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Denver, Colorado, USA: ACL, pp. 89–98.

Song, Z., Biesa, A., Strassel, S., Ellis, J., Mitamura, T., Dang, H., Yamakawa, Y., and Holm, S. 2016. Event nugget and event coreference annotation. In *Proceedings of the 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, NAACL HLT 2016.

Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P.C., Erickson, B., Miller, T., Lin, C., Savova, G., and Pustejovsky, J. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* **2**(2014): 143–54.

Sun, W., Rumshisky, A., and Uzuner, O. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*. **20**(5): 806–13. Oxford University Press.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of SemEval 2013*, Atlanta, Georgia, USA: ACL, pp. 1–9.

Van Den Akker, C., Aroyo, L., Cybulska, A., Van Erp, M., Gorgels, P., Hollink, L., Jager, C., Legene, S., van der Meij, L., Oomen, J., van Ossenbruggen, J., Schreiber, G., Segers, R., Vossen, P., and Wielinga, B. 2010. Historical event-based access to museum collections. In *Proceedings of EVENTS2010*, Athens, Greece: ACL.

van Son, C., van Erp, M., Fokkens, A., and Vossen, P. 2014. Hope and fear: interpreting perspectives by integrating sentiment and event factuality. In *Proceedings of LREC 2014*, Reykjavik, Iceland, European Language Resources Association (ELRA), pp. 26–31.

Vendler, Z. 1967. Verbs and times. In *Linguistics and Philosophy*, pp. 97–121. Ithaca, NY: Cornell University Press.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of SemEval-2007*, Prague, Czech Republic: ACL, pp. 75–80.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden: ACL, pp. 57–62.

Vossen, P., Rigau, G., Serafini, L., Stouten, P., Irving, F., and Van Hage, W. 2014. NewsReader: recording history from daily news streams. In *Proceedings of LREC 2014*, Reykjavik, Iceland, European Language Resources Association (ELRA).

Yaghoobzadeh, J., Ghassem-Sani, G., Mirroshandel, S. A., and Eshaghzadeh, M. 2012. ISO-TimeML event extraction in Persian text. In *Proceedings of COLING 2012*, Mumbai, India, pp. 2931–2944.

Zavarella, V., and Tanev, H. 2013. FSS-TimEx for TempEval-3: extracting temporal information from text. In *Proceedings of SemEval 2013*, Atlanta, Georgia, USA, ACL, pp. 58–63.