



UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

EVENT DETECTION AND
CLASSIFICATION FOR THE DIGITAL
HUMANITIES

Rachele Sprugnoli

Advisor

Dr. Sara Tonelli

Università degli Studi di Trento

April 2018

Abstract

In recent years, event processing has become an active area of research in the Natural Language Processing community but resources and automatic systems developed so far have mainly addressed contemporary texts. However, the recognition and elaboration of events is a crucial step when dealing with historical texts: research in this domain can lead to the development of methodologies and tools that can assist historians in enhancing their work and can have an impact both in the fields of Natural Language Processing and Digital Humanities. Our work aims at shedding light on the complex concept of events adopting an interdisciplinary perspective. More specifically, theoretical and practical investigations are carried out on the specific topic of event detection and classification in historical texts by developing and releasing new annotation guidelines, new resources and new models for automatic annotation.

Keywords

Natural Language Processing; Event Detection; Event Classification; Digital Humanities; Information Extraction.

Contents

Acknowledgements	1
1 Introduction	3
1.1 Motivations and Goals	3
1.2 Contributions	8
1.3 Structure of the Thesis	9
2 Event Definition, Detection and Processing in NLP	11
2.1 Events in Linguistics	12
2.2 The IE Perspective on Events	14
2.2.1 First studies on Events	14
2.2.2 Evaluation Campaigns	16
2.2.3 Comparison of Event Annotation Schemes	21
2.2.4 The Italian Case: EVENTI evaluation exercise	22
2.3 Crowdsourcing for Event Annotation	33
2.3.1 Recognition of Events from Raw Text	35
2.4 Event Processing Beyond the News Domain	39
2.5 Chapter Summary	46
3 Event Definition, Detection and Processing in DH	47
3.1 Semantic Web Event Ontologies	48
3.1.1 Semantic Event Models	49

3.1.2	Modelization of Biographical Events	51
3.2	Projects	54
3.2.1	FDR/Pearl Harbor	55
3.2.2	Semantics of History	56
3.2.3	Agora	57
3.2.4	DIVE+	58
3.2.5	NanoHistory	60
3.2.6	Bringing Lives to Light: Biography in Context . . .	61
3.2.7	BiographyNet	63
3.2.8	Semantic Biographies Based on Linked Data	64
3.3	Weak Points of Current Efforts in DH	66
3.4	Chapter Summary	69
4	Survey on Event Definition and Annotation	71
4.1	What is an Event in History?	72
4.1.1	Questionnaire Description	73
4.1.2	Questionnaire Results	74
4.1.3	Discussion	84
4.2	Chapter Summary	85
5	Events in Historical Texts: Guidelines and Manual Corpus Annotation	87
5.1	Event Annotation Guidelines	88
5.1.1	Event Linguistic Realization	88
5.1.2	Event Extent	90
5.1.3	Semantic Classes	93
5.2	Dataset Construction	104
5.2.1	Corpus Description	104
5.2.2	Corpus Annotation	107
5.2.3	Interlinking Events and Content Types	112

5.3	Chapter Summary	115
6	Events in Historical Texts: Automatic Annotation	119
6.1	Data Preparation	120
6.2	CRF Classifiers	121
6.2.1	Evaluation	122
6.3	Bi-LSTM Approach	125
6.3.1	Evaluation	126
6.4	Systems Comparison and Discussion	135
6.5	Chapter Summary	139
7	Conclusions	141
7.1	Answers to Research Questions	142
7.2	Future Directions	144
	List of Publications	147
	Bibliography	153
	Appendices	191
A	Case Study: the Shoah Ontology and the Events of Movement	193
A.1	Tracing Movements of Italian Shoah Victims	193
A.1.1	Related Work	195
A.1.2	Workflow	196
A.1.3	Quantitative and Qualitative Data Analysis	200
A.2	Appendix Summary	203
B	Questionnaire: What is an Event in History?	205
B.1	English Questionnaire	205

B.1.1	Introduction	205
B.1.2	Part 1 - Events in historical texts	206
B.1.3	Part 2 - Natural Language Processing	208
B.2	Italian Questionnaire	211
B.2.1	Parte 1 - Eventi nei testi storici	212

List of Tables

2.1	Annotated events, temporal expressions, signals and temporal relations in the EVENTI corpus.	29
2.2	Average number of annotated events, temporal expressions and temporal relations per 1,000 tokens in the EVENTI corpus.	29
2.3	The <i>De Gasperi corpus</i> : quantitative data	30
2.4	Results, in terms of F1, of the best automatic systems on both contemporary news and the De Gasperi corpus together with the drop in performances between the two domains.	32
2.5	Crowd <i>vs.</i> Expert: Event token annotation in English (EN) and Italian (IT)	38
2.6	Corpora including event annotation in the news domain.	44
2.7	Corpora including event annotation in different domains other than news.	45
3.1	Summary of the main characteristics of DH projects described in Section 3.2.	68

4.1	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.	76
4.2	Italian 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.	77
5.1	Statistics on the <i>Histo Corpus</i>	107
5.2	Annotated events per class and text genre together with the total amount of annotations. The asterisk indicates whether the class has a statistically significant difference in the distribution over the two genres.	111
5.3	Number (#) of annotated CT in the <i>Histo Corpus</i> and Cohen’s kappa (k), calculated between two annotators, for each type of CT. An asterisk marks that in all the cases there is a statistically significant differences in the distribution of CTs over the two genres.	115
6.1	Performance, in terms of precision (P), recall (R) and F1, of the CRF model for event mention detection with different settings of features.	122
6.2	Performance, in terms of precision (P), recall (R) and F1, of the CRF model for the event detection+classification task.	123

6.3	Performance of the CRF classifier on event mention detection with different context windows.	124
6.4	Performance of the CRF classifier on event mention classification with different context windows.	124
6.5	Average precision (P), recall (R) and F1 over three runs of the BiLSTM system with the configuration suggested by Reimers and Gurevych [2017a].	127
6.6	Results of the BiLSTM system with different optimization algorithms on the event mention detection only task. . . .	129
6.7	Results of the BiLSTM system with different optimization algorithms on the event detection+classification task. . . .	129
6.8	Performance with different character embeddings options on the event mention detection only task.	130
6.9	Performance with different character embeddings options on the task of event detection+classification.	130
6.10	Precision, Recall and F1 score with the CRF and Softmax classifiers in the event mention detection only task.	131
6.11	Precision, Recall and F1 score with the CRF and Softmax classifiers in the event detection+classification task.	131
6.12	Results obtained with different pre-trained word embeddings for the event mention detection only task.	133
6.13	Results obtained with different pre-trained word embeddings for the event detection+classification task.	133
6.14	Results of the CRF classifier and the BiLSTM model with the best configuration for the event mention detection only task.	136
6.15	Results of the CRF classifier and the BiLSTM model with the best configuration for the task including both the detection and the classification of event mentions.	136

A.1	Division of victims on the basis of the <code>death_description</code> property.	201
A.2	Number of events dated and georeferenced in the <i>LOD Navigator</i> together with the five most frequent locations for each event.	202

List of Figures

1.1	Timeline of evaluation campaigns (above) and workshops (below) in the field of event detection and processing. The “Time & Space Track @Semeval 2015” includes the Time-Line, QA TempEval and Clinical TempEval tasks. The De-RiVE workshop series focuses on the detection, representation, and exploitation of events in the Semantic Web field, while EVENTS2017 is about the modeling of events in Cultural Heritage.	5
2.1	Taxonomy of eventualities, image taken from [Dölling et al., 2014].	13
2.2	Comparison of different event annotations. Red squares highlight event triggers while 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.	21
2.3	Corpus creation cycle: from the printed book to the annotated documents.	28
2.4	Distribution (in percent) of event classes in the EVENTI corpus.	31
2.5	Distribution (in percent) of temporal relations in the EVENTI corpus.	32

2.6	Instructions for the event and temporal expression detection task from raw text (English case).	37
3.1	The <code>Temporal Entity</code> hierarchy in CIDOC CRM.	50
3.2	Core classes of BIO, a vocabulary for biographical information.	52
3.3	Two subclasses of the class <code>:Ecclesiastical_Event</code> in Bio CRM with their corresponding allowed roles. Example taken from [Tuominen, 2016].	54
3.4	Event recognition and role assignment in the FDR/Pearl Harbor project. Figure taken from [Ide and Woolner, 2004].	56
3.5	SEM instance filled with information: the event name <i>Independence of Indonesia</i> is associated with an actor <i>Soekarno</i> and a place <i>Surabaya</i> . Image taken from [Van Erp Marieke and Schreiber., 2011].	58
3.6	Validation and annotation of event identification in texts and video in the DIVE+ project.	59
3.7	Manual modeling of events in the Nanohistory project. . .	60
3.8	Geo-visualization of the chronology of events in Emma Goldman’s life. Screenshot from http://metadata.berkeley.edu/emma/	63
3.9	Event visualization in the Person perspective of WarSampo portal at https://www.sotasampo.fi/en/	65
4.1	Answers to the question “In which country do you live?”. Belgium, Canada, Greece and Republic of Serbia are the countries gathered in the <i>Other</i> category.	74

4.2	Answers to the question “How would you define your field of research?”. Respondents could give more than one preference. Under the <i>Other</i> category, 11 additional fields were indicated: i.e., digital humanities, archaeology, history of books, history of religions, epistemology of history, spatial history, legal history, methodology of the historical research, history of Christianity, history of publishing, history of the East.	75
4.3	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.	81
5.1	Mapping of our HISTO classes (left) to the second-level HTOED categories (right). Image created with RAWGraphs [Mauri et al., 2017].	94
6.1	Example of a file in the CAT XML format (left) and in the corresponding converted BIO/IOB2 notation (right) for the two tasks.	120
6.2	Architecture of the BiLSTM network with a CRF-classifier adapted from [Reimers and Gurevych, 2017a].	126
6.3	Comparison of F1 scores for each evaluated event class. The CLOTHES class is not in the Figure because it was not present in the test set.	137
A.1	Properties whose content was extracted from CDEC LOD dataset.	197
A.2	Example of output of the Wiki Machine.	199

Acknowledgements

First and foremost I am deeply grateful to my supervisor Sara Tonelli, who has always believed in my skills even when I doubted myself, and whose advice has always been very precious to me.

I would like to thank the reviewers and members of my thesis examination committee for their insightful comments and constructive feedback: Elisabetta Jezek, Thierry Declerck and Bernardo Magnini. Thanks also to Anna Feltracco for her help with inter-annotator agreement, to Giulia Marchesini for the collaboration on the annotation of Content Types, and to Charles Callaway for proofreading some sections of this thesis.

I owe my gratitude to Amedeo Cappelli, who many years ago offered me the opportunity to work in Trento, and to Tommaso Caselli, who introduced me to the world of temporal processing and who, over time, has become a great co-author and a reference point.

I gratefully acknowledge all my colleagues in Trento. It is not easy to live so far from home, but I was lucky to find great company in my work environment. A special thank you goes to Manuela Speranza, who welcomed me in FBK 13 years ago when I did not know anyone in Trento, and to Luisa Bentivogli, Danilo Giampiccolo and all the CELCT crew: we've been through a lot together and the memory of Emanuele Pianta will unite us forever.

A big thank you to my officemates in the DH group: Giovanni Moretti, Stefano Menini and Alessio Palmero Aprosio. In particular, I am highly

indebted to Giovanni, the most talented developer I've ever known but also a supportive colleague, a generous friend, and a person I can always trust.

A deep thank goes to all my friends - Cristina, Elena, Francesca, Manuela, Marta, Medea, Paola, Serena C., Serena G., Silvia and Vittoria: we do not see each other often, but I know that I can always count on you.

Finally, my utmost gratitude goes to my family, to the members who are no longer with us but whose example has formed me, and to those that brighten up my life everyday, without ever making me feel lonely. In particular, I'm grateful to my grandmother Libia and my aunt Luana, who firmly believed in the importance of the emancipation of women through education, my supportive uncle Armando, my brothers-in-law, my caring and protective sisters, my nephews and niece who are a constant joy for me, my generous, loving and tenacious mom and my devoted dad, a profoundly honest man and a model for my career.

Chapter 1

Introduction

We are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline.

Karl Popper [1963]

The aim of this work is to advance the research in event detection and classification by adopting an interdisciplinary perspective. More specifically, we draw on knowledge from the fields of Information Extraction and Digital Humanities to address the problem of event detection and classification in historical texts, opening up a field of inquiry so far underestimated in the area of Temporal Information Processing.

This introductory Chapter is meant to delineate the motivations and goals of our research, highlight its main contributions, and provide an overview of how the thesis is organized.

1.1 Motivations and Goals

In the last 25 years, several systems performing event extraction have been presented within the Natural Language Processing (NLP) community. Diverse approaches aimed at building timelines from large document collections have been implemented, and technologies to support automatic sto-

1.1. MOTIVATIONS AND GOALS

rytelling have become a relevant research topic [Ashish et al., 2006, Jung et al., 2011, Laparra et al., 2015, Vossen et al., 2015]. In addition, event processing has been addressed from a variety of perspectives, from data visualization (see, for example, [Fulda et al., 2016]) to knowledge representation and modelling [Allen and Ferguson, 1994]. 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 over the years.

In NLP, an important distinction related to event recognition and processing concerns two different research areas: in the field of Topic Detection and Tracking (TDT), the identification of events is assimilated to 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 thesis, we focus on the latter perspective, since it has led to several standardisation 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 TDT 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 [Atefeh and Khreich, 2013].

The timeline² in Figure 1.1, 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 event detection and processing organized starting from the third Message Understanding Conference (MUC-3).

¹According to the Linguistic Data Consortium (LDC, <https://www ldc.upenn.edu/>) annotation guidelines of the TDT task, “a topic is defined as an event or activity, along with all directly related events and activities” [Strassel, 2005].

²An interactive version of the timeline is available online: http://dhl lab.fbk.eu/Timeline_events/.

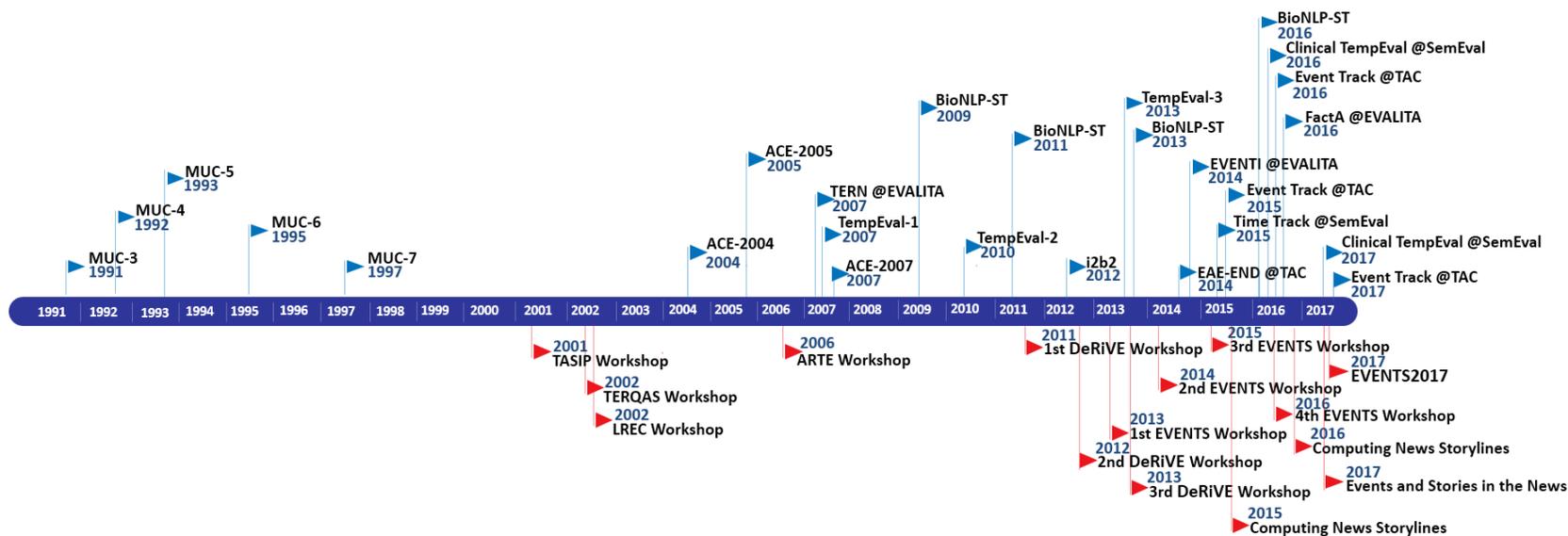


Figure 1.1: Timeline of evaluation campaigns (above) and workshops (below) in the field of event detection and processing. The “Time & Space Track @Semeval 2015” includes the TimeLine, QA TempEval and Clinical TempEval tasks. The DeRiVE workshop series focuses on the detection, representation, and exploitation of events in the Semantic Web field, while EVENTS2017 is about the modeling of events in Cultural Heritage.

1.1. MOTIVATIONS AND GOALS

The high concentration of initiatives in the last years makes evident the increasing interest in automatic analysis and processing of temporal information, especially in the IE community. The workshops and the evaluation campaigns shown in the timeline mainly focused on the analysis of contemporary news articles because the newswire domain has been the most extensively investigated one in IE. More recently, much attention has been devoted also to the temporal processing of social media text, clinical records and bio-medicine documents leading to the development of domain-specific event definitions [Intxaurreondo et al., 2015, Bethard et al., 2015b, Björne and Salakoski, 2011].

The notion of event has been studied also in Humanities and Social Sciences disciplines, which the NLP community has hardly taken into account. In particular, the recognition and elaboration of events is a crucial step when dealing with history-related matters. Even if in the contemporary historiographical approach, History is no more considered a mere chronological accumulation of events in a coherent timeline [Bloch, 1954], events are still the building blocks of historical knowledge with which historians construct their system of ideas about the past [Oakeshott, 2015, Shaw, 2010].

We do not enter in the details of the philosophical debate that opposes events as concrete individual things in the world [Davidson, 2001] to events seen as products of narrative language [Mink, 1978], a debate still ongoing among historians³. However, IE assumes a neo-Davidsonian approach [Higginbotham, 2000, Parsons, 1990] according to which events are predicates that can take various textual forms (called *mentions*) corresponding to different parts of speech: as such, events can be recognised and detected within texts.

³For a summary of the debate seen from the point of view of historical practice, please refer to [Shaw, 2013].

1.1. MOTIVATIONS AND GOALS

Given these premises, this thesis is built around three main themes: the notion of event in IE, the notion of event in History, and the cross-fertilization between these two perspectives so to satisfactorily deal with the task of event detection and processing in historical texts. These themes led us to define two main research questions:

Research Question 1. How can the notions of event in IE and History be combined?

Research Question 2. How can methods and techniques of Information Extraction be applied to the recognition and classification of events in historical texts in a way that it satisfies the actual needs of domain experts?

The first question will be addressed with a comprehensive study and a critical analysis of the state of the art in event definition, detection and processing but also with the employment of elicitation techniques to involve domain experts. In particular, we will review initiatives, projects and approaches in IE and in the Digital Humanities, a field of research in which traditional humanities and computational methods encounter, interact and support each other⁴.

To answer the second question we will follow the traditional multi-stage process to train and test an IE algorithm [Pustejovsky and Stubbs, 2012]: we will thus develop annotation guidelines and an annotated corpus and we will build our own models for the automatic detection and classification of event mentions in historical texts.

⁴Despite its roots date back to the work of Padre Busa on the Index Thomisticus started at the end of the '40s, the expression “Digital Humanities” is still much discussed and there is no consensus about a precise definition. For a detailed examination of this definitional issue, please refer, among others, to [Nyhan Julianne and Vanhoutte, 2013, Lunenfeld et al., 2012, Klein and Gold, 2016]

1.2 Contributions

The contributions of this thesis are both theoretical and practical.

From the theoretical point of view, innovative aspects lay in our effort of integrating knowledge coming from two different areas so to create a holistic, interdisciplinary view of the object of our research, i.e. the notion of event. This approach led us to develop:

- an exhaustive and original survey of the state of the art in event definition and processing in both NLP and Digital Humanities (DH);
- the first investigation on how events are defined by historians in their everyday research practice supported by a comparison with ongoing standardization efforts in the NLP community.

From the practical point of view, the contributions are the following:

- co-organization of the first evaluation exercise on Temporal Information Processing on Italian texts and introduction, for the first time, of historical texts in an evaluation exercise on Temporal Information Processing;
- design and implementation of experiments on the possibility of adopting crowdsourcing techniques for the annotation of events;
- development of new annotation guidelines for the detection and classification of event mentions specifically designed for historical texts;
- creation of a novel corpus of historical texts annotated with events made available to the research community;
- release of word embeddings pre-trained on a corpus of historical texts and of models for the automatic annotation of events developed using

1.3. STRUCTURE OF THE THESIS

two distinctive approaches: traditional linear-chain Conditional Random Fields (CRFs) from one hand and a neural architecture from the other.

Annotated corpus, pre-trained historical embeddings and best models are available on GitHub: <https://github.com/dhfbk/Histo>.

1.3 Structure of the Thesis

This thesis is organized as follows. In Chapter 2 we provide a detailed overview of how events have been defined in NLP with a focus on the efforts, undertaken in the area of Information Extraction in the last decades, that were directed towards their automatic detection and processing. Within this area, we present our works for the promotion of Temporal Information Processing in Italian and our studies on the application of crowdsourcing techniques on event annotation. A different perspective is introduced in Chapter 3 in which we describe and analyse works and projects carried out on event definition, detection and processing in the DH field. In Chapter 4 we report on an investigation we performed involving historians in an 'event definition and annotation' exercise through a web questionnaire. Chapter 5 presents our guidelines for event mention annotation in historical texts, based on the outcome of this investigation. We also give details on a newly created corpus annotated following the guidelines. Chapter 6 describes our experiments aimed at the development of an automatic system for event detection and classification: both a Conditional random fields classifier and a neural architecture are tested and their results compared. Chapter 7 provides a summary of the thesis and discusses the lesson learned from this research work. The thesis is completed by two appendices. The first one is dedicated to a collaborative work we carried out on the extraction of events of movement of Italian Shoah victims from a Linked Open Dataset.

1.3. STRUCTURE OF THE THESIS

The second Appendix reports the content of the questionnaire discussed in Chapter 4.

Chapter 2

Event Definition, Detection and Processing in NLP

In this chapter, after a brief introduction on the role of events in linguistics (Section 2.1), we provide an overview of the way events have been defined in Information Extraction (IE) (Section 2.2), with a focus on the first seminal works in this field (Section 2.2.1) and on the different evaluation campaigns organized over the years (Section 2.2.2). Event annotation schemes presented within these campaigns are compared in Section 2.2.3 while we illustrate the evaluation exercise on temporal processing for the Italian language in Section 2.2.4. We also account for multilingual event processing, presenting tasks and corpora that cover languages other than English, for the use of crowdsourcing in event annotation (Section 2.3) and for new domains involved in recent event definition efforts (Section 2.4).

Part of the Chapter is based on our publication on “Natural Language Engineering” journal [Sprugnoli and Tonelli, 2017]. The description of the EVENTI evaluation exercise, co-organized during the PhD in the context of EVALITA 2014¹, in Section 2.2.4 has already been presented in [Caselli et al., 2014a]. The language specific adaptation of TimeML annotation scheme to Italian and the creation of Ita-TimeBank, that is at the basis

¹<http://www.evalita.it/2014>

2.1. EVENTS IN LINGUISTICS

of EVENTI, are described in a book chapter published during the PhD and written in collaboration with Tommaso Caselli [Caselli and Sprugnoli, 2017]. Within EVENTI, the annotation of historical texts, presented in [Speranza and Sprugnoli, IN PRESS], was part of the De Gasperi Project, described in [Sprugnoli et al., 2016]. The methodology and the results of the crowdsourcing experiments reported in Section 2.3 are published in [Sprugnoli and Lenci, 2014] and [Caselli et al., 2016].

2.1 Events in Linguistics

Although, at an intuitive level, 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 that Sasse [2002] defines “so horribly confusing”. 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]². In particular, nominals exhibit a strong semantic ambiguity due to polysemy, showing alternations between eventive and non-eventive readings [Apresjan, 1974, Pustejovsky, 2005, Melloni and Others, 2011]: for example, *administration* denotes an event in *spending grew during his ad-*

²“on peut s’attendre à voir le verbe et le nom comme deux poles (...), constituer une sorte de champ magnétique où les catégories oscillent en subissant l’attraction soit de l’un soit de l’autre, soit des deux” trad. *we can expect to see the verb and the noun acting as poles constitute a kind of magnetic field where the categories fluctuate as they are attracted either one or the other, or both.*

2.1. EVENTS IN LINGUISTICS

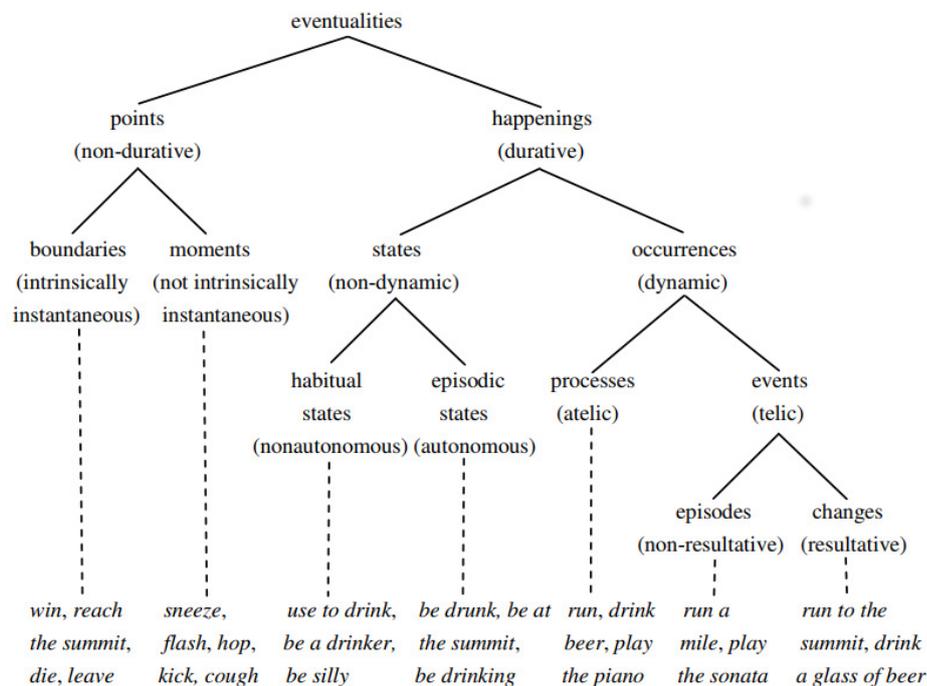


Figure 2.1: Taxonomy of eventualities, image taken from [Dölling et al., 2014].

ministration and a human group in *this administration is doing well*. A crowdsourcing experiment on the polysemy of Italian nominal events is reported in Section 2.3.

In linguistics, the best-known classification of events is the one proposed by [Vendler, 1957], 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, 1991a]. Moreover, in the literature, all types of actions, states and processes often fall under the cover term “eventualities”, coined by [Bach, 2008] in his work on the algebra of

2.2. THE IE PERSPECTIVE ON EVENTS

events and re-elaborated in [Dölling et al., 2014].

2.2 The IE Perspective on Events

In Information Extraction, 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*). 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 reported in the Introduction (see Figure 1.1) gives a graphical overview of these initiatives; below we instead provide a detailed description and analysis of works and evaluation campaigns related to the notion of event in the field of IE.

2.2.1 First studies on Events

In 2001, during the Workshop “Temporal and Spatial Information Processing” (TASIP), 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. Their system achieves a precision of 0.55 and a recall of 0.60 but no baseline is reported. Schilder and Habel [Schilder and Habel, 2001] present a tool for the automatic annotation of temporal expressions in German news: they include both time-denoting expressions, like dates, and event-denoting expressions. The latter are de-

2.2. THE IE PERSPECTIVE ON EVENTS

defined as expressions that have an implicit time dimension and are either verbs or noun phrases, especially nominalisations. The list of markable nouns is limited to those directly connected to a temporal expression or a temporal preposition (e.g., *after the takeover 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 with a cascade of Finite State Transducers combined with an ontology containing event-denoting nouns in the financial domain and information on event types [Schilder and Habel, 2003]. The recognition of time-denoting and event-denoting expressions is evaluated in a single run showing a precision of 0.98 and a recall of 0.97 on nominal chunks and a precision of 0.95 and a recall of 0.94 on verbal chunks. In the only extraction on nominal event descriptions, however, precision is 0.66 and recall is 0.44. Given that the authors have used only simple features for this task, they consider these results as a baseline for further experiments. Finally, Katz and Arosio [Katz and Arosio, 2001] propose a method to annotate temporal relations at sentence level, limiting events to verbs. Their event identifier system achieves a precision of 0.56 and a recall of 0.61 but no baseline results are provided. The three works, even if focusing on different languages (i.e., English and German) and datasets (i.e, news on an earthquake in Afghanistan, economic articles, and random sentences from the British National Corpus³), highlight the need to achieve a consensus on a definition of event, aimed also at making automatic approaches comparable.

In that same year, Setzer [Setzer, 2001] presents STAG (*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 instantana-

³<http://www.natcorp.ox.ac.uk/>

2.2. THE IE PERSPECTIVE ON EVENTS

neous 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 STAG, TimeML [Pustejovsky et al., 2003] 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, SemAf/Time Working Group, 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 Portuguese TimeBank [Costa and Branco, 2012] and the Romanian TimeBank [Forascu and Tufi, 2012] .

2.2.2 Evaluation Campaigns

Parallel to the works reported in the previous Section, several evaluation campaigns on temporal information extraction and processing have been carried out. As shown in Figure 1.1, such campaigns have become very frequent in the last decade, with some years characterized by multiple evaluations. Each evaluation exercise resulted in the definition of a specific annotation scheme and in the development of annotated corpora. This Section provides an overview of the main annotation efforts and campaigns.

2.2. THE IE PERSPECTIVE ON EVENTS

International Evaluation Campaigns

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 to the ones achieved in other MUC tasks such as Named Entity Recognition (NER) 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], while the best systems in the NER and CO tasks achieved an F-score of 0.93 and 0.62 respectively [Miller et al., 1998, Gaizauskas et al., 1995]. 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 to fill in some of these slots.

In the “Event Detection and Recognition” task, run for three years in the context of the ACE (Automatic Content Extraction) program, 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

2.2. THE IE PERSPECTIVE ON EVENTS

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 have 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, 48% of the events in the training corpus of ACE2005 belong 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 program [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 ontology⁴ (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

⁴The full ACE event ontology is reported in the annotation guidelines of the evaluation campaign [Linguistic Data Consortium, 2005].

2.2. THE IE PERSPECTIVE ON EVENTS

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 coreference.

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 (EAE) and a pilot task on event nugget detection (END) [Song et al., 2016] in 2014,⁵ and these same tasks are included also in the Event Track of TAC KBP 2015, 2016 and 2017.⁶

Although the TAC KBP 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.

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 focusing 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 con-

⁵<http://www.nist.gov/tac/2014/KBP/Event/index.html>

⁶<http://www.nist.gov/tac/2015/KBP/Event/index.html>

2.2. THE IE PERSPECTIVE ON EVENTS

sisted of 6 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 et al., 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., part-of-speech) 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 10 points, with F1-scores all below 0.72.

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.⁷ The TimeLine task addressed coreference 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., 2015a] being reconfirmed as task in other two SemEval editions, in 2016 and 2017 respectively. For more information about event processing in the clinical domain, see Section 2.4.

⁷<http://alt.qcri.org/semeval2015/>

2.2. THE IE PERSPECTIVE ON EVENTS

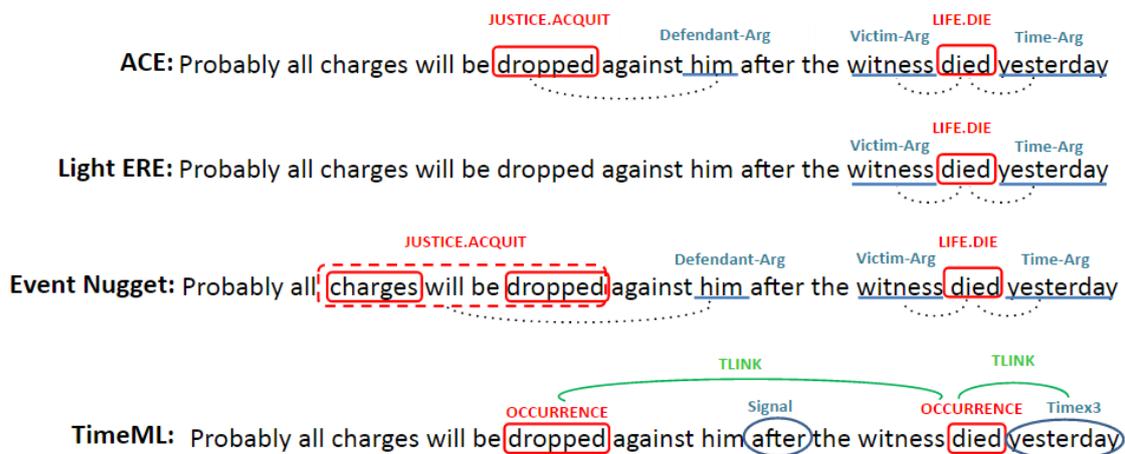


Figure 2.2: Comparison of different event annotations. Red squares highlight event triggers while blue underlinings identify other annotated elements that in ACE, Light ERE and Event Nugget constitute event arguments. Connections between events and arguments are displays in dotted lines. For TimeML, temporal links are in green.

2.2.3 Comparison of Event Annotation Schemes

As a wrap-up of the different annotation schemes described in this section, we present in Figure 2.2 the same sentence annotated according to ACE, Light ERE, Event Nugget, and TimeML guidelines. Differences in event types among ACE, Light ERE and Event Nugget are minimal (in this example are even null), while there is more variation concerning extent. ACE, Light ERE and TimeML annotate only events as single tokens, while Event Nugget schema annotates 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

2.2. THE IE PERSPECTIVE ON EVENTS

ERE and Event Nugget combine information on events with their argument structure, while in TimeML the temporal dimension acquires more relevance, having its roots in Allen’s interval algebra [Allen, 1990].

2.2.4 The Italian Case: EVENTI evaluation exercise

TempEval-2 has boosted multilingual research in Temporal Processing by making TimeML-compliant data sets available in six languages, including Italian. Unfortunately, partly due to the limited size of that corpus (less than 30,000 tokens), no system was developed for Italian. As a consequence, there was no complete system for Temporal Processing for the Italian language, but only independent modules for event [Robaldo et al., 2011, Caselli et al., 2011b] and temporal expressions processing (Heidel-Time) [Strötgen et al., 2014]. To fill this gap, the EVENTI evaluation exercise⁸ was built upon previous TempEval evaluation campaigns to promote research in Temporal Processing for Italian by offering a complete set of tasks for comprehension of temporal information in written text. The task was co-organized during the PhD in collaboration with Manuela Speranza (FBK) and Tommaso Caselli (University of Groningen). Task details and results were published in the proceedings of the EVALITA 2014 workshop [Caselli et al., 2014a].

EVENTI Annotation

The EVENTI exercise is based on the EVENTI annotation guidelines, a simplified version of the Italian TimeML Annotation Guidelines⁹ (henceforth, It-TimeML) [Caselli and Sprugnoli, 2015], using four It-TimeML tags: TIMEX3, EVENT, SIGNAL and TLINK (Temporal LINKS). For

⁸<https://sites.google.com/site/eventievalita2014/>

⁹We have created a website dedicated to the IT-TimeML initiative containing data, guidelines and publications: <https://sites.google.com/site/ittimeml/>

2.2. THE IE PERSPECTIVE ON EVENTS

clarity’s sake, we report only the changes which have been applied to It-TimeML.

- The **EVENT** tag is used to annotate all mentions of events including verbs, nouns, prepositional phrases and adjectives. With respect to TimeML, we have introduced exceptions to the minimal chunk rule for multi-token event expressions, namely collocations and idiomatic expressions contained in reference resources, mainly the Italian Dictionary by Tullio de Mauro [De Mauro, 2000] (the list of multi-token expressions created for this purpose is available online¹⁰). We have simplified the annotation of events realized by adjectives and prepositional phrases by restricting it to the cases in which they occur in predicate position with the explicit presence of a copula or a copular verb. For example, in *L’assemblea è stata solidale con loro* / “The meeting was in solidarity with them both the copula” *è* and the adjective *solidale* are annotated as events. **EVENTs** are categorized on the basis of 7 classes: **REPORTING** (events describing declarations, narrations, or the provision of information about something, e.g. *dire* / “to say”, *dichiarazione* / “declaration”), **PERCEPTION** (events describing physical perception, e.g.: *vedere* / “to see”, *sentire* / “to hear”), **ASPECTUAL** (events providing information on a particular phase or aspect of an event, such as the beginning or the end, e.g. *iniziare* / “to start”, *finire* / “to finish”), **I_ACTION** (intensional actions, such as *provare* / “to try”, *offrire* / “to offer”), **I_STATE** (intensional states, such as *credere* / “to believe”, *avere paura* / “be afraid”, and the modal verbs), **STATE** (circumstances in which something obtains or holds true, such as *presenza* / “presence” in *La presenza dei nostri ministri*), **OCCURRENCE** (all other types of events, e.g. *crescere* / “grow”,

¹⁰<https://sites.google.com/site/eventievalita2014/data-tools/poliremEVENTI.txt>

2.2. THE IE PERSPECTIVE ON EVENTS

uragano / “hurricane”). In the sentence below for example, we identify two events, i.e. *guerra* / “war” and *finita* / “finished”, respectively of class OCCURRENCE and ASPECTUAL.

Invece la guerra non è ancora finita / “On the contrary, war is not finished yet”

- The TIMEX3 tag is used for the annotation of temporal expressions. Temporal expressions can be composed of a single token or a sequence of tokens with a temporal meaning and include the following classes: DATE (e.g. 15/01/2015), TIME (e.g. 11:00), DURATION (e.g. *due mesi* / “two months”), and SET, i.e. a frequency of occurrence (e.g. *ogni anno* / “every year”). No changes have been made with respect to It-TimeML thus the creation of empty, non-text consuming TIMEX3 tags whenever a temporal expressions can be inferred from a text-consuming one is allowed. In the example below, 1949 and 1945 are annotated as DATE, in addition a non-text consuming TIMEX3 tag expressing the duration of 5 years is created.

La guerra è durata dal 1940 al 1945 / “the war lasted from 1940 to 1945”.

- SIGNAL is used to annotate function words indicating how events and temporal expressions are temporally related to each other (e.g. *dopo* / after). In EVENTI, we have annotated only SIGNALs indicating temporal relations. In the following example, *quando* / when is a SIGNAL indicating the presence of a temporal relation between the events *fuggito* / “dashed off” and *avviata* / “headed”).

Quando il treno è fuggito, la massa di popolo, lentamente, silenziosamente, si è avviata attraverso le uscite / “When the

2.2. THE IE PERSPECTIVE ON EVENTS

train dashed off, the mass of people, slowly, silently, headed through the exits”

- TLINK identifies a temporal relation between two temporal expressions, two events, or a temporal expression and an event (e.g. “The Police ARRESTED a suspect AFTER a CHASE”). More specifically, following TimeML we distinguished the following classes of TLINKs: INCLUDES (one event/timex includes the other), IS_INCLUDED (the inverse of INCLUDES), BEFORE (an event/timex occurs before another), AFTER (the inverse of BEFORE), BEGINS (a timex or an event marks the beginning of another timex or event), BEGUN_BY (the inverse of BEGINS), ENDS (a timex or an event marks the ending of another event or timex), ENDED_BY (the inverse of ENDS), IBEFORE, i.e. immediately before (one event/timex occurs immediately before the other), IAFTER, i.e. immediately after (the inverse of IBEFORE), SIMULTANEOUS (two events happen at the same time), MEASURE (used to link an event and a duration which provides information on how long the event last), and IDENTITY (links the elements of causative constructions, light verb constructions and copular constructions). The TLINK tag did not undergo any changes in terms of use and attribute values. Major changes concern the definition of the set of temporal elements that can be involved in a temporal relation. Details on this aspect are reported in the description of subtask C in Section 2.2.4. In the example below, the TLINK between *rispondevano* / “answered” and *ringraziando* / “thanking” is of class SIMULTANEOUS (in the sentence above, on the other hand, we have a BEFORE TLINK between *fuggito* and *avviata*).

I poveri giovani rispondevano ringraziando / “The poor young people answered by thanking them”

2.2. THE IE PERSPECTIVE ON EVENTS

EVENTI Subtasks

The EVENTI evaluation exercise is composed of a Main Task and a Pilot Task. The difference between these tasks lays in the type of data to be processed: contemporary news in the former, historical news in the latter. Each task consists of a set of subtasks in line with previous TempEval campaigns and their annotation methodology.

The subtasks proposed are:

- Subtask A: determine the extent, the type and the value of temporal expressions in a text according to the TIMEX3 tag definition. For the first time, empty TIMEX3 tags were taken into account in the evaluation;
- Subtask B: determine the extent and the class of the events in a text according to the EVENT tag definition;
- Subtask C: identify temporal relations in raw text. This subtask involves performing subtasks A and B and subsequently identifying the pairs of elements (event - event and event - timex pairs) which stand in a temporal relation (TLINK) and classifying the temporal relation itself. Given that EVENTI is an initial evaluation exercise in Italian and to avoid the difficulties of full temporal processing, we have further restricted this subtask by limiting the set of candidate pairs to: i) pairs of main events in the same sentence; ii) pairs of main event and subordinate event in the same sentence; and iii) event - timex pairs in the same sentence. All temporal relation values in It-TimeML are used.
- Subtask D: determine the value of the temporal relation given two gold temporal elements (i.e. the source and the target of the relation) as

2.2. THE IE PERSPECTIVE ON EVENTS

defined in Task C (main event - main event; main event - subordinate event; event - timex).

Data

The EVENTI evaluation exercise is based on the EVENTI corpus, which consists of 3 datasets: the Main task training data, the Main task test data and the Pilot task test data.

The news stories distributed for the Main task are taken from the Ita-TimeBank [Caselli et al., 2011a]. Two expert annotators have conducted a manual revision of the annotations for the Main task to solve inconsistencies mainly focusing on harmonizing event class and temporal relation values. The annotation revision has been performed using CAT¹¹ [Bartalesi Lenzi et al., 2012], a general-purpose web-based text annotation tool that provides an XML-based stand-off format as output. The final size of the EVENTI corpus for the Main task is 130,279 tokens, divided in 103,593 tokens for training and 26,686 for test.

The Main task training data have been released to participants in two separate batches¹² through the Meta-Share platform¹³. Annotated data are available under the Creative Commons Licence Attribution-NonCommercial-ShareAlike 3.0 to facilitate re-use and distribution for research purposes.

The Pilot test data, henceforth the *De Gasperi Corpus*, is made by ten articles written by Alcide De Gasperi, a prominent Italian politician at both the national and international level. De Gasperi was the first prime minister of the Italian Republic after the end of the monarchy and is considered one of the founding fathers of the European Union. In the first period of his long career, which lasted more than 50 years, he was a journalist at the newspaper *Il Trentino*. We selected the documents for the creation

¹¹<http://dh.fbk.eu/resources/cat-content-annotation-tool>

¹²ILC Training Set: <http://goo.gl/3kPJkM>; FBK Training Set: <http://goo.gl/YnQWm1>

¹³<http://www.meta-share.eu/>

2.2. THE IE PERSPECTIVE ON EVENTS

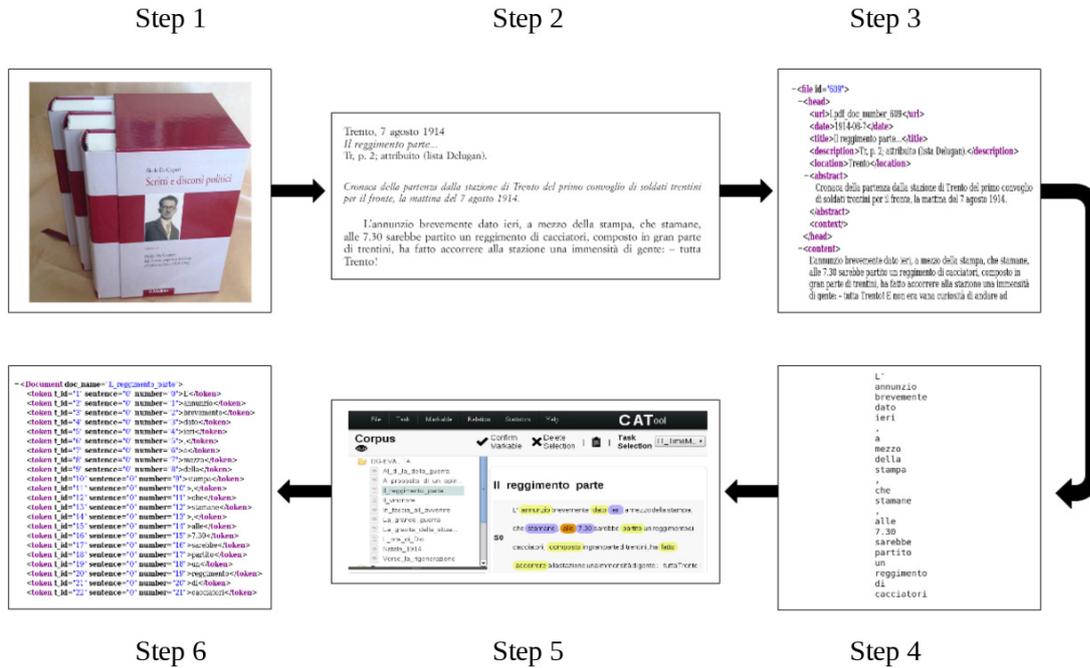


Figure 2.3: Corpus creation cycle: from the printed book to the annotated documents.

of our annotated corpus from a critical edition of De Gasperi’s writings published in 2006 [De Gasperi, 2006] (step 1 in Figure 2.3); in particular, we focused on articles published in *Il Trentino* in 1914 and related to the outbreak of World War I. The process for the creation and annotation of our corpus started from the files kindly provided to us by the publisher of De Gasperi, et al. (2006) in PDF format (step 2). Then we converted the PDF files into XML format: the resulting XML files contain the text of the document together with some metadata information, such as the title and the date of the original publication (step 3). We used TextPro, a suite of NLP tools for Italian (Pianta et al., 2008), to automatically tokenize each text and split it into sentences (step 4). In this format, we were able to upload the documents into the CAT annotation tool. An expert annotator performed the manual annotation following the specifications defined for the EVENTI task (step 5). CAT stores the annotated files in a stand-off XML format that is flexible and easy to manipulate (step 6). Both the

2.2. THE IE PERSPECTIVE ON EVENTS

original text files and these annotated XML files are freely available for research purposes¹⁴.

	Main Training	Main Test	Pilot Test /De Gasperi Corpus
EVENTs	17,835	3,798	1,195
TIMEX3s	2,735	624	97
SIGNALs	932	231	62
TLINKs	3,500	1,061	382

Table 2.1: Annotated events, temporal expressions, signals and temporal relations in the EVENTI corpus.

	Main Training	Main Test	Pilot Test /De Gasperi Corpus
EVENTs	172.1	142.4	239
TIMEX3s	26.4	23.3	19.0
TLINKs	33.7	39.7	76.4

Table 2.2: Average number of annotated events, temporal expressions and temporal relations per 1,000 tokens in the EVENTI corpus.

Table 2.1 reports the total number of each annotated element type in the Main task training set, in the Main task test set, and in the Pilot test set. The comparison between the average number of EVENTs, TIMEX3s and TLINKs annotated in the three datasets is given Table 2.2 while Table 2.3 shows the distribution of the annotations into the different classes in the De Gasperi Corpus. The Pilot corpus clearly shows a higher density of events (238 vs. 172.1 and 142.4 for training and test, respectively) and temporal relations (76.4 vs. 33.7 and 39.7 for training and test, respectively). On the other hand, the average number of temporal expressions in the two corpora is comparable, however closer analysis highlighted an important difference: the 54% of temporal expressions in the De Gasperi corpus is fuzzy (e.g. *i sacrifici dell'⟨ora presente⟩*) or non-specific (e.g. *nei ⟨giorni⟩ del dolore*).

¹⁴<http://dh.fbk.eu/technologies/eventi-datasets-temporal-information-processing-italian>

2.2. THE IE PERSPECTIVE ON EVENTS

EVENT		TLINK	
OCCURRENCE	623 (52.13%)	SIMULTANEOUS	111 (29.06%)
STATE	277 (23.18%)	BEFORE	85 (22.25%)
I.STATE	146 (12.22%)	IDENTITY	69 (18.06%)
I.ACTION	75 (6.28%)	IS_INCLUDED	67 (17.54%)
ASPECTUAL	31 (2.59%)	INCLUDES	18 (4.71%)
REPORTING	29 (2.43%)	AFTER	18 (4.71%)
PERCEPTION	14 (1.17%)	ENDED_BY	4 (1.05%)
		MEASURE	3 (0.79%)
		BEGUN_BY	2 (0.52%)
		ENDS	2 (0.52%)
		IBEFORE	2 (0.52%)
		BEGINS	1 (0.27%)
		IAFTER	0 (0%)

TIMEX3	
DATE	65 (67.01%)
TIME	18 (18.56%)
DURATION	11 (11.34%)
SET	3 (3.09%)

Table 2.3: The *De Gasperi corpus*: quantitative data

We illustrate in Figure 2.5 the distribution of the class values of **EVENTs** and the distribution of the temporal values for **TLINKs** in the three datasets. The most frequent classes are **OCCURRENCE** and **STATE**, followed by **I.STATE** and **I.ACTION**. The high prevalence of occurrences and states is not surprising as these classes encode the objects of a narrative (e.g. contemporary news or historical texts) or what people “speak about”. On the other hand, more interesting results are provided by the relatively high presence of the **I.STATE** and **I.ACTION** classes. According to the TimeML definitions, these classes are used either to express intensional relations or speculations about “possible worlds” between events. They are markers of subjectivity along the axis of event factivity, pointing out that people do not limit themselves to “speak about” happenings but they also speculate on these happenings. The higher frequency of **I.STATE** in the Pilot corpus with respect to the Main datasets is due to the fact that the Pilot dataset is mainly composed of editorial comments which frequently contain perspectives on and speculations about the world. Additional evidence is also

2.2. THE IE PERSPECTIVE ON EVENTS

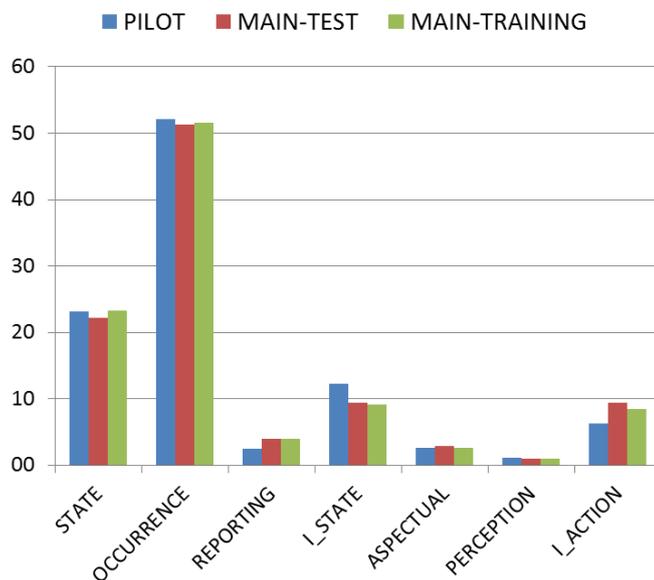


Figure 2.4: Distribution (in percent) of event classes in the EVENTI corpus.

the lower frequency of the REPORTING class in the Pilot dataset than in the Main task. The high presence of personal opinions influences also the temporal structure of the texts, whereby most events are not ordered chronologically but presented as belonging to the same time frame on top of which the De Gasperi expresses his opinions and suggests future and alternative courses of events. As a matter of fact, the most frequent temporal relation in the Pilot task is SIMULTANEOUS. On the other hand, in the Main task there is an evident preference for IS_INCLUDED. The Main task is composed of news articles, where events tend to be more often linked to temporal containers (e.g. temporal expressions or other events) to facilitate understanding of stories by readers.

Results

Table 2.4 reports the results obtained by the best system in the Main task and the results obtained by the best system in the Pilot task for the recognition and classification of TIMEX3s, EVENTS, and TLINKs. If we

2.2. THE IE PERSPECTIVE ON EVENTS

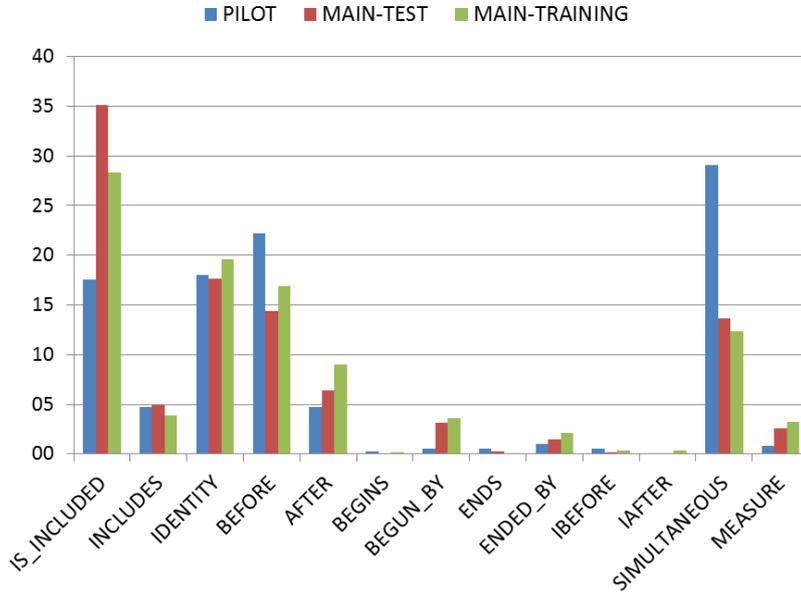


Figure 2.5: Distribution (in percent) of temporal relations in the EVENTI corpus.

	Main task (contemporary news)	Pilot task (De Gasperi)	Performance drop
TIMEX3	0.8	0.678	0.122 (15%)
EVENT	0.671	0.604	0.067 (10%)
TLINK	0.264	0.185	0.079 (30%)

Table 2.4: Results, in terms of F1, of the best automatic systems on both contemporary news and the De Gasperi corpus together with the drop in performances between the two domains.

compare the results, we notice that the drop in performance affects all three evaluated elements.

The proportionally greatest drop concerns the annotation of TLINKs (where the loss amounts to 30%), while for EVENTS we only go from F1=0.671 to F1=0.604 (with a difference of slightly less than 10%). By observing the corpora used for the evaluation, we might suppose that there is a correlation between the performance obtained by participant systems and the distribution of the different categories of TLINKs and EVENTS in the two corpora. In fact, EVENTS of type OCCURRENCE are by far the most

2.3. CROWDSOURCING FOR EVENT ANNOTATION

frequent in both corpora (representing in both cases more than 50% of the total, followed by EVENTS of type **STATE**). In the case of TLINKs, on the other hand, there is more variability in the distribution, with **SIMULTANEOUS** being the most frequent category in the De Gasperi corpus (almost 30% of the total) and **IS_INCLUDED** the most frequent category in the corpus of contemporary news (almost 35% of the total). Although the overall results of the EVENTI pilot task on historical texts might appear encouraging, we must observe that they do not take into consideration the issue of whether a system able to produce the (IT)TimeML-based output required by this task is actually useful for the everyday work of historians, as there has been no adaptation of the (IT)TimeML annotation guidelines to the history domain.

2.3 Crowdsourcing for Event Annotation

The evaluation campaigns mentioned in the previous Section have relied on a substantial amount of manual effort for data annotation and checking and, as a consequence, with considerable investment in terms of time and costs. In this Section, we instead report on experiments we conducted to study the feasibility of using crowdsourcing for the annotation of events.

Crowdsourcing, i.e. participative online activities in which the voluntary undertaking of a task is proposed to a group of individuals of varying knowledge [Estellés-Arolas and González-Ladrón-de Guevara, 2012], has been extensively used for lots of tasks in NLP [Wang et al., 2013]. In particular, crowdsourcing platforms such as Mechanical Turk and Crowd-Flower have been used for the the creation of language resources and the annotation of text, images and speech; see, among others, [Chamberlain et al., 2016, Kovashka et al., 2016, Sprugnoli et al., 2017b]. However, the use of crowdsourcing in the perspective of Temporal Processing has been

2.3. CROWDSOURCING FOR EVENT ANNOTATION

mainly limited to studies which aim at assessing the difficulty of the task and the salience of linguistic and extralinguistic cues with a particular focus on the temporal relations rather than on all the subtasks involved [Mani and Schiffman, 2005, Caselli and Prodanof, 2010, Ng and Kan, 2012]. As for event detection, in Aroyo and Welty [2012] the focus of crowdsourcing is not on assessing the ability of the crowd to perform that specific task, but on disagreement as a “natural state” suggesting that event semantics is imprecise and varied.

A first experiment, initially proposed in our Master thesis [Sprugnoli, 2012] and then further extended during the PhD [Sprugnoli and Lenci, 2014], evaluated the ability of the crowd in detecting event nominals in Italian, pointing out the complexity of this task due to the presence of ambiguous patterns of polysemy [Jezek, 2008]. The same task was performed by two expert annotators¹⁵ so to allow a comparison of the results. Moreover, a gold standard was created by combining expert annotations and performing a reconciliation on disagreements.

The accuracy of the results, calculated on the gold standard and obtained with the crowd experiment (74%), proved not to be comparable to the accuracy obtained by experts before reconciliation (93%). The inter-coder agreement confirms the problematic nature of this task for non-expert contributors that obtained a kappa of 0.34 whereas experts achieved an agreement of 0.81. These results shows that the recognition of nominal events is not an intuitive task, easily accomplished using only practical instructions made available to non-expert contributors.

In the following subsection, another set of experiments, conducted in collaboration with Dutch scholars, is described: further information are published in an LREC 2016 paper [Caselli et al., 2016].

¹⁵The expert annotators involved in this experiment were the author of the thesis and another Italian native speaker with proven knowledge of Italian linguistics and previous experience in the field of semantic annotation.

2.3. CROWDSOURCING FOR EVENT ANNOTATION

2.3.1 Recognition of Events from Raw Text

The study described above has been extended going beyond the only nominal events and taking into consideration two languages, i.e. English and Italian. More specifically, following the works by Soberon et al. [Soberon et al., 2013] and Inel et al. [Inel et al., 2013], we designed a set of experiments in English and in Italian where the crowd is asked to identify event descriptions and temporal expressions, and, then, on top of these crowd annotated elements, the presence of temporal relations and their values. In other words, we asked the crowd to perform the Temporal Processing task from raw texts. The aim is to replicate a more realistic annotation scenario as the crowd workers perform all subtasks involved in the temporal annotation of documents from raw text data. To this end, we extracted 200 random sentences from the English and Italian TimeBank corpora [Pustejovsky et al., 2002, Caselli et al., 2011a] and we adopted the CrowdTruth metrics [Inel et al., 2014] for cleaning the data from spammers and evaluating their quality. The reuse of texts from TimeBank corpora already annotated with temporal information allowed us to make comparisons between crowd and expert annotations. Event detection and temporal expression identification were merged in a single task but this subsection reports only on the identification of event descriptions, since this is the focus of the PhD work.

The CrowdTruth Metric

The goal of the CrowdTruth methodology is i) to distinguish between high-quality and low-quality workers, and ii) to assess how well a given label is expressed by the input data [Soberon et al., 2013, Aroyo and Welty, 2014]. The first step is to transform workers judgments into annotation vectors so to take advantage of cosine similarity measures. The length of

2.3. CROWDSOURCING FOR EVENT ANNOTATION

the vector depends on the number of possible answers in a question, while the number of such vectors depends on the number of questions contained in the task. If the worker selects a particular answer, its corresponding component would be marked with 1, and 0 otherwise.

In the specific case of our experiment, we build two vectors for each annotated unit having the dimension equal to the total number of words in the sentence and the option “none”, if no word in the sentence refers to an event. We then compute a media unit vector by adding up all the workers annotation vectors for that unit. Next, we apply two worker metrics, computed using the cosine similarity, to understand how close each worker performs compared to the others on a single unit and how much a worker disagrees with the rest of the workers taking into consideration all the units. If the worker values are below a given threshold, the worker is marked as low-quality and his/her annotations are discarded.

To determine how well an annotation is expressed in a unit we compute the unit-annotation score, or clarity score, on the spam-filtered data. This metric is measured for each possible annotation on each unit as the cosine between the unit vector for that annotation and the media unit vector.

For event detection we ran an overall of seven different jobs, i.e., 3 jobs for English data and 4 jobs for Italian data. Workers were allowed to select both single tokens and multi-token expressions and then had to decide if the identified word(s) was an event. For each sentence, we collected a total of 15 judgments. Each worker was allowed to annotate a maximum of 10 sentences (e.g. 10 judgments). Instructions for English and Italian had the same content: in particular, we used a basic definition of event as “something that has happened, is happening or will happen in the future”. An example of the English instructions and of the settings of the task is illustrated in Figure 2.6.

As for the English data, a total of 372 workers from USA, UK, Australia

2.3. CROWDSOURCING FOR EVENT ANNOTATION

and Canada participated in the experiments; 124 (33.33%) were identified as spammers on the basis of CrowdTruth metrics. For the Italian dataset, we collected judgments from 371 workers from Italy. By applying the CrowdTruth metrics, we identified 115 spammers (30.99%). We further analyzed the data with the clarity score to compare the ability of the crowd(s) versus the experts: the higher is the clarity score, the more accurate and reliable are the crowd judgments.

The screenshot shows a web interface titled "Highlight Events & Temporal Expressions In Text". It includes an "Instructions" section with the following text: "Read carefully the TEXT below and highlight the WORDS or PHRASES that refer to events or are temporal expressions." Below this are two bullet points: "EVENT is a word/phrase that refers to something that has happened, is currently happening or will/may happen in the future (e.g. meeting, concert, eating, World Cup 2015, 5th Annual Congress, Battle of Waterloo, etc.)" and "TEMPORAL EXPRESSION is a word/phrase which refers to different expressions of time (e.g. yesterday, 2001, at noon, from 15.00 p.m. to 16.00 p.m., every year, v)". Below the instructions is a text box containing the sentence: "Pastor James Allmen of the fellowship church and school in Ashburn has led the anti-Saudi campaign." The words "led" and "anti-Saudi campaign" are highlighted in green. Below the text box are two steps: "STEP 1: In the text above, HIGHLIGHT the words/phrases that refer to an EVENT or are TEMPORAL EXPRESSIONS." and "STEP 2: Indicate the type of each HIGHLIGHTED word/phrase (EVENT or TEMPORAL EXPRESSION)". Below step 2 are three dropdown menus: "led" with "Event" selected, "anti-Saudi campaign" with "Event" selected, and "campaign" with "Event" selected. Each dropdown menu has a "[x]" button next to it.

Figure 2.6: Instructions for the event and temporal expression detection task from raw text (English case).

Event Detection

As for English, 1,296 tokens were judged as expressing an event, while in Italian only 1,040 tokens were annotated. To compare the performance of the crowd(s) and the experts for this subtask, we analyzed the number of overlapping tokens per clarity score thresholds. In Table 2.5 we report, for different clarity thresholds and, for both languages, the number of to-

2.3. CROWDSOURCING FOR EVENT ANNOTATION

kens marked as events by the crowd(s) together with the overlap with the experts.

CLARITY	# CROWD EVENT TOKENS		CROWD-EXPERT OVERLAPPING EVENT TOKENS	
	EN	IT	EN	IT
	≥ 0.2	1121	566	355 (31.66%)
≥ 0.3	628	358	270 (42.99%)	251 (70.11%)
≥ 0.4	314	184	168 (53.50%)	145 (78.80%)
≥ 0.5	164	100	103 (62.80%)	80 (80%)
≥ 0.6	71	60	52 (73.23%)	51 (85%)

Table 2.5: Crowd *vs.* Expert: Event token annotation in English (EN) and Italian (IT)

With no threshold for clarity score, we identified 444 tokens (34.26%) which overlap expert annotation in the English data (TimeBank corpus), covering 84.25% of all events annotated by experts (527). On the other hand, for the Italian data, we identified 473 tokens which overlap with expert annotation (Ita-TimeBank corpus), covering only 53.87% of all event tokens annotated by the experts (878). With different clarity thresholds the number of annotated tokens by the crowd(s) get reduced (e.g. from 1,121 tokens with score ≥ 0.2 to 71 tokens with score ≥ 0.6 for English; from 566 tokens with score ≥ 0.2 to 60 tokens with score ≥ 0.6 for Italian) but the quality of the annotation improves, i.e., they are more reliable and in line with the expert data.

By analyzing mismatches in the event annotation between the crowd(s) and the experts we can observe that:

- with a threshold ≥ 0.3 , 274 tokens in English are candidates of multi-token events such as noun phrases (*national callup*, *global embargo*), phrasal verbs (*fall apart*, *going up*), multiword expressions (*coup d'état*),

2.4. EVENT PROCESSING BEYOND THE NEWS DOMAIN

verbs accompanied by auxiliaries (*were offset, have fallen*), and copular constructions (*were lower*). As for Italian, with the same threshold, 77 tokens are possible multi-token events: noun phrases (*raccolta diretta, sconfitta definitiva*), verbs accompanied by auxiliaries (*ha commentato*), multiword expressions (*5,000 metri*), and proper nouns (*Cross della Vallagarina*);

- the crowd annotation in English has identified 12 candidate event tokens which are missing in the expert data and has also provided annotations for 4 sentences which the experts did not annotate. The missing annotations are mainly nominal events (*trading, operations*) or verbs (*clobbered, cut*). Similarly, the crowd annotation for Italian has identified 13 event tokens missing in the expert data. The missing annotations mainly correspond to named events (*Flushing Meadows*) and nominal events (*scadenza, cadute*).

What emerges from this experiment is that English and Italian crowd annotations of events have commonalities in terms of the text span of the markables and errors. For instance, for events the crowd tends to prefer larger textual span annotations than the experts by including participants (e.g. *held the stronghold Police arrested six Protestants*) or even assuming complex event representations (e.g. *driving under the influence of alcohol*). Differences in the text span of events should not be considered as real errors in the annotations but signal a more holistic understanding of events from the crowd(s) with respect to the analytic models preferred by the experts, in line with the results in [Aroyo and Welty, 2012].

2.4 Event Processing Beyond the News Domain

Most evaluation exercises presented so far were concerned with event processing in the news domain. Only recently, NLP researchers have started

2.4. EVENT PROCESSING BEYOND THE NEWS DOMAIN

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.¹⁶ Following the guidelines¹⁷, 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 et al., 2013] and of the Clinical TempEval evaluation, organized within SemEval 2015, 2016 and 2017 and aimed at assessing the performance of temporal information extraction systems on clinical notes and pathology reports.¹⁸ One of the subtask of Clinical TempEval concerns the identification of the textual span of event descriptions and the assignment of values to a set of event attributes, i.e. modality, degree, polarity and semantic type. In the first two years of the evaluation exercise, the best system achieved equal or even higher results with respect of human agreement. BluLab, by using Support Vector Machines (SVM) algorithms and linguistic features, had a F-score of 0.87 on span identification and was always above 0.82 on the assignment of the different attributes values [Bethard et al., 2015a, Velupillai et al., 2015]. In 2016, also UTHealth adopted SVM but added more features, embeddings and information from domain specific dictionaries achieving an F-score of 0.93 on span identification and above 0.85 on attribute assignment [Bethard et al., 2016, Lee et al., 2016].

¹⁶The University of Colorado at Boulder has recently proposed an extension of the THYME guidelines integrating ISO-TimeML, the Stanford Event coreference [Lee et al., 2012] and the CMU Event coreference guidelines [Hovy et al., 2013] under the name of Richer Event Description (RED). RED adopts the TimeML wide definition of events and annotates events, temporal expressions and entities, as well as temporal, coreference and casual relations [Ikuta et al., 2014]. For more information, see the guidelines: <https://github.com/timjogorman/RicherEventDescription>.

¹⁷<http://clear.colorado.edu/compsem/documents/>

¹⁸2015: <http://alt.qcri.org/semeval2015/task6/>; 2016: <http://alt.qcri.org/semeval2016/task12/>; 2017: <http://alt.qcri.org/semeval2017/task12/>

2.4. EVENT PROCESSING BEYOND THE NEWS DOMAIN

In 2017, the focus of ClinicalTempEval has shifted towards domain adaptation: systems were trained on a clinical condition (colon cancer data) and tested on another clinical condition (brain cancer data) using an unsupervised approach and also a supervised one but with a limited quantity of training in-domain data. The best system, LIMSI-COT, proposed a deep learning approach for event detection using Long Short-Term Memory Networks (LSTMs) and a linear SVM for each attribute. The system obtained an F-score of around 0.70 in the unsupervised setting for both span identification and attribute assignment and around 0.75 in the supervised one showing a consistent drop in the performances with respect to the previous year [Bethard et al., 2017, Tourille et al., 2017].

Since 2009, several editions of the BioNLP shared task 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 SVM and linguistic features, while the third one is rule-based [Hakala and Landeghem, 2013, Björne and Salakoski, 2013, Xuan Quang Pham, Minh Quang Le, 2013]. SVM is the machine learning model used also by the best systems in the 2016 task edi-

2.4. EVENT PROCESSING BEYOND THE NEWS DOMAIN

tion [Li et al., 2016, Lever and Jones, 2016]. A deep learning approach has been proposed too, being ranked second in the identification of localization events of bacteria [Mehryary et al., 2016].

Event extraction from **social media** is another emerging area of research [Atefeh and Khreich, 2013]. Most of the works in this field address the task as a clustering problem following the TDT approach, for example using an unsupervised method and focusing on the detection of unspecified new events [Edouard et al., 2017, Zhou et al., 2017]. Other works deal with the retrieval of retrospective events in microblogs, such as Twitter: among others, Metzler et al. [2012] propose a temporal query expansion technique to retrieve a ranked list of event summaries, having the events classified in different categories and types. Ritter et al. [2012] test a different approach applying 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.

In order to account for all corpora annotated so far with event information in different domains and languages, we report a summary in Tables 2.6 and 2.7. The information presented in the tables was gathered through the direct analysis of the resources downloaded from the Web and merging data from scientific papers. Resources listed in the tables have been annotated following different schemes and cover five domains: corpora in the news domain are reported in Table 2.6 while other domains are covered in 2.7 among which history, a domain discussed in the Chapter 4. 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. The number of corpora in the list shows the interest of the NLP

2.4. EVENT PROCESSING BEYOND THE NEWS DOMAIN

community in event processing. The most recent corpora confirm the trend towards new domains, new languages and more complex tasks integrating event extraction.

2.4. EVENT PROCESSING BEYOND THE NEWS DOMAIN

Domain	Corpus	Lang	#Tokens	#Files	#Events
NEWS	ACE 2005 (training) ^a	EN	259,889	599	4167
		ZH	307,991	633	3332
	French TimeBank [Amsili et al., 2011]	FR	15,423	109	2,115
	Romanian TimeBank [Forascu and Tufi, 2012]	RO	65,375	181	7,926
	TimeBankPT [Costa and Branco, 2012]	PT	69,702	182	7,887
	Persian TimeBank [Yaghoobzadeh et al., 2012]	FA	26,949	43	4,237
	Catalan TimeBank 1.0 ^b	CA	75,800	210	12,342
	Spanish TimeBank 1.0 ^c	ES	75,800	210	12,641
	BCCWJ-TimeBank [Asahara, 2013]	JA	56,518	54	3,824
	EVENTI corpus [Caselli et al., 2014b]	IT	130,279	366	21,633
	TempEval 1 (training) ^d	EN	52,740	162	5,150
		ZH	32,788	61	1,204
	TempEval 2 (training+test) ^e	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 ^f	EN	102,375	276	12,534
	FactBank [Saurí and Pustejovsky, 2009]	EN	77,000	208	9,500
	EventCorefBank (ECB) [Lee et al., 2012]	EN	-	482	2,533
	ECB+ [Cybulska and Vossen, 2014]	EN	377,367	982	15,003
	Light ERE ^g [Mott et al., 2016]	ZH ^h	127,458	171	481
		EN	101,191	171	369
	Rich ERE ^g [Mott et al., 2016]	ZH ^h	127,458	171	1,491
		ES	101,191	171	2,933
	Event Nugget [Mitamura et al., 2015]	EN	336,126	351	10,719
	TimeLine [Minard et al., 2015]	EN	29,893	90	915
	MEANTIME ⁱ [Erp et al., 2008]	EN	13,981	120	2,096
		IT	15,676	120	2,208
		ES	15,843	120	2,223
		NL	14,647	120	2,223

^a <https://catalog.ldc.upenn.edu/LDC2006T06>

^b <https://catalog.ldc.upenn.edu/LDC2012T10>

^c <https://catalog.ldc.upenn.edu/LDC2012T12>

^d <http://www.timeml.org/tempeval/>

^e <http://timeml.org/tempeval2/>

^f <http://www.cs.york.ac.uk/semeval-2013/task1>

^g Light ERE, Rich ERE and Event Nugget corpora include both news and discussion forum data

^h Number of characters instead of the number of tokens

ⁱ <http://www.newsreader-project.eu/results/data/wikinews/>

^j <https://catalog.ldc.upenn.edu/LDC2012T01>

Table 2.6: Corpora including event annotation in the news domain.

2.4. EVENT PROCESSING BEYOND THE NEWS DOMAIN

Domain	Corpus	Lang	#Tokens	#Files	#Events
CLINICAL	i2b2 [Sun et al., 2013]	EN	178,000	349	30,000
	Clinical TempEval (Train+Dev) [Bethard et al., 2015a]	EN	533,393	440	59,864
BIOMEDICAL	GENIA [Kim et al., 2008]	EN	-	1,000	36,114
SOCIAL MEDIA	Twitter NLP [Ritter et al.]	EN	19,484	1,000	-
HISTORY	ModeS TimeBank ^j	ES	25,611	102	1,261
	De Gasperi Corpus [Caselli et al., 2014b]	IT	5,671	10	1,195

^j <https://catalog.ldc.upenn.edu/LDC2012T01>

Table 2.7: Corpora including event annotation in different domains other than news.

2.5 Chapter Summary

In this Chapter we provide an overview of past and current trends in the definition, automatic detection and processing of events in the field of NLP. Details on evaluation campaigns and annotation efforts are given with a special focus on the work we have done in the context of the EVENTI evaluation exercise and on the results of some crowdsourcing experiments conducted during the PhD. Beside the news domain, the most investigated one in the field of NLP, we give an account of approaches adopted in other domains, such as social media and biomedicine showing that a careful adaptation of existing annotation schemes is necessary to apply the outcome of past and current research activities to new domains.

The next Chapter is dedicated to the analysis of event definition and processing in the field of Digital Humanities, thus another domain is taken into account. Ontologies and projects related to events and developed in this field are described and shortcomings of current approaches highlighted.

Chapter 3

Event Definition, Detection and Processing in DH

A number of works in the past have tried to capture and analyse the semantics of texts in the Humanities using a combination of Semantic Web technologies and NLP approaches [Meroño-Peñuela et al., 2015]. In this context, particular attention has been devoted to the representation, detection and processing of events seen as crucial elements to be analysed for a deep understanding of textual sources. Starting from this observation, in Section 3.1 we describe ontologies created and used in the Digital Humanities (DH) area to model events while in Section 3.2 we report on projects conducted in the field of DH in which language technologies were employed to automatically extract events, which were then modeled and represented using ontologies and RDF statements. A specific Section is then devoted to an overview of shortcomings we detected in current approaches to event detection and processing in DH.

3.1 Semantic Web Event Ontologies

The design and development of knowledge organization systems, such as thesauri and ontologies¹, have gained much attention in the Digital Humanities. For example, the Historical Thesaurus of the Oxford English Dictionary (HTOED) is a comprehensive historical thesaurus that organizes the meaning of all the words contained in the Oxford English Dictionary with a diachronic approach [Kay et al., 2009a] (for more information, please see Section 5.1.3).

In this Section we focus on event-centric ontologies: in recent years, several RDF (Resource Description Framework)/OWL (Web Ontology Language) ontologies have been proposed to model events and their relationships with different kinds of entities with the goal of improving the access to Cultural Heritage collections, organizing information in historical archives, representing the semantics of bibliographic records and of biographical texts.

In the following, we first provide a brief description of the most used ontologies in the field of DH: in general, all these ontologies want to model factual aspects of events by representing at least four fundamental dimensions, that is *what*, *where*, *when* and *who*. Later, in Section 3.1.2, we concentrate on how life events are modeled in biographies. We chose to focus on biographical information as a specific example of data characterized by the presence of specific types of events, such as birth, death, carrier achievements, relocations. Moreover, biographies have recently gained much attention with several interdisciplinary projects (see Sections 3.2.6, 3.2.7 and 3.2.8) and dedicated conferences (i.e., “Biographical Data in a

¹When writing about ontologies, we follow the definition by Struder et al. [1998] thus an ontology is “a formal, explicit specification of a shared conceptualization”. In this view, adopted in particular in the computer science community, an ontology is a computational artifact expressed in a machine readable format [Guarino et al., 2009].

3.1. SEMANTIC WEB EVENT ONTOLOGIES

Digital World in 2015 and 2017”).

3.1.1 Semantic Event Models

In the **Event Ontology**², initially developed in the area of music research, an event is defined as “an arbitrary classification of a space/time region, by a cognitive agent.” An event is characterized by its participating agents and can also have sub-events, factors, products, temporal and spatial dimensions. However, no way to model roles, views, and the temporary validity of properties is provided. Different domain ontologies have been designed to be used in conjunction with the Event Ontology, for example the Music Ontology³ to model music-related information such as festivals and song writing.

LODE (Linking Open Descriptions of Events)⁴ defines only one class, i.e. **event**: its focus is explicitly on publishing records on events as reported in news or by a historian and on enhancing the interoperability with other vocabularies and ontologies [Shaw et al., 2009]. In addition to properties defining place, time, involved objects or agents of events, LODE includes a property that links media objects to the events they illustrate. For this reason, LODE has been especially adopted in media projects, see [Khrouf and Troncy, 2016] among others.

Differently from LODE that is conceived to provide a minimal modeling of events, the **CIDOC Conceptual Reference Model** (CIDOC CRM)⁵ is quite vast having, in its last version, 89 classes and 151 properties [Le Boeuf et al., 2017]. In particular, CIDOC CRM covers events related to the Cultural Heritage domain, e.g. the acquisition and transfer of an artwork, the creation and the copying of a text [Doerr, 2003]. Figure

²<http://motools.sourceforge.net/event/event.html>

³<http://musicontology.com/>

⁴<http://linkedevents.org/ontology/>

⁵<http://www.cidoc-crm.org/>

3.1. SEMANTIC WEB EVENT ONTOLOGIES

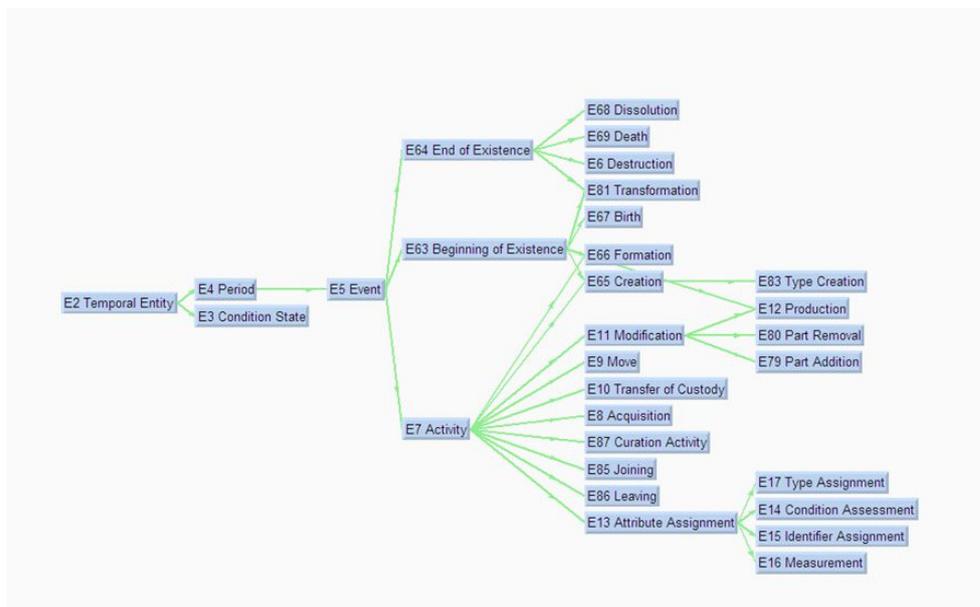


Figure 3.1: The Temporal Entity hierarchy in CIDOC CRM.

3.1⁶ displays the hierarchy of the Temporal Entity class which includes all temporal phenomena, i.e. events and states, happening over a limited period of time. The success of CIDOC CRM is especially due to the fact that it is an ISO standard (ISO 21127:2006) and easy to extend to meet the complexity of specific types of data: for example, extensions have been designed to describe archaeological processes and techniques [Meghini et al., 2017] or are under development to provide foundational support to historical research projects as in the “Data for History” initiative⁷.

Cultural Heritage is also the domain of the EDM (Europeana Data Model) ontology whose focus is on the interoperability with other ontologies and standards so to allow an easy ingestion of cultural objects from different institutions within Europeana⁸, the EU digital platform for cultural heritage. One of EDM main classes is **event**, defined as a change of

⁶Image taken from “The CIDOC CRM, a Standard for the Integration of Cultural Information”, presentation by Stephen Stead, <http://slideplayer.com/slide/6045743/>

⁷<http://dataforhistory.org/>

⁸<https://www.europeana.eu/>

3.1. SEMANTIC WEB EVENT ONTOLOGIES

state in the cultural, social or physical system or a coherent phenomenon or cultural manifestation happening in a specific time and location [The European Union, 2012].

SEM (Simple Event Model) has four main classes: **Event**, **Actor** (i.e. event participant), **Place** and **Time** [Van Hage et al., 2011]. Interactions between the instances, identified through URIs, are represented with RDF triples however, no constraints are put on the RDF vocabularies to be adopted. Generalization and aggregation of events are supported, moreover relations of any kind between events and other instances can be modeled with the `sem:eventProperty` relation. In addition, other ontologies and schema can be combined with SEM, making it very flexible and easily adaptable to different domains. Particular attention is devoted in the representation of the provenance of a statement and in modeling the presence of different views on the same description of an event. Thanks to its characteristics, SEM has been used in several projects both in the NLP (see, among others, the NewsReader project [Fokkens-Zwirello et al., 2013, Vossen et al., 2016]) and in the DH field (see Section 3.2).

3.1.2 Modelization of Biographical Events

Biographies are the subject of a wide range of studies in the Digital Humanities as demonstrated by the establishment of a dedicated biennial conference series called “Biographical Data in a Digital World”⁹: several repositories of biographies are available in digital format and specific schemata have been proposed to model life events, so to improve the analysis and understanding of these repositories.

BIO¹⁰ describes a person’s life seen as a series of interlinked events.

⁹Website of 2015 conference: <http://www.biographynet.nl/biographical-data-in-a-digital-world/>.
Website of 2017 conference: <https://sites.google.com/view/bd2017/>.

¹⁰<http://vocab.org/bio/>

3.1. SEMANTIC WEB EVENT ONTOLOGIES

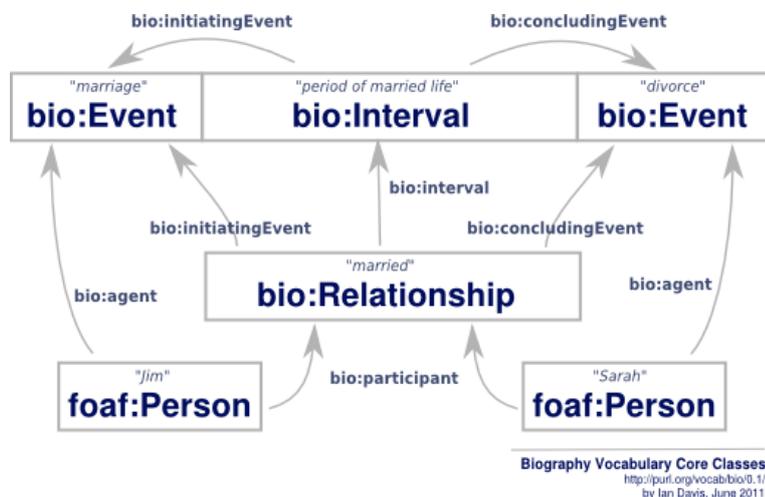


Figure 3.2: Core classes of BIO, a vocabulary for biographical information.

The vocabulary, expressed in OWL, has four core classes as shown in Figure 3.2¹¹: **Person**, **Event**, **Relationship** and **Interval**. As for the **Event** class, BIO proposes a framework of 37 event types: some of these types applies to all people (e.g., **Birth**, **Death**), others are more specific (e.g., **Coronation**, **BarMitzvah**). Each event is characterized by four properties: **Date**, **Place**, **State** (i.e., territory involved in an event), and **Position** (i.e, employment position or public office). Other properties are used to relate an event to an agent (e.g., **Employer**, **Officiator**) or to temporally order an event with respect to another event (e.g. **Following Event**, **Preceding Event**). An extension of BIO has been proposed within the **Shoah Ontology**, a domain ontology that formally describes concepts and relationships characterizing the life and persecution of Jews in Italy between 1943 and 1945 [Brazzo and Mazzini, 2015]. Here, the ontology class called **Persecution** is used to represent all main events related to the persecution of the victims (arrest, detention, deportation to a Nazi camp, transfer to another camp, liberation, death in a massacre). This class is connected to the **Person** class that is based on BIO extended with

¹¹Image taken from <http://vocab.org/bio/>

3.1. SEMANTIC WEB EVENT ONTOLOGIES

additional anagrafic/genealogical properties (e.g. `niece_nephewOf`). The application *LOD Navigator*, designed during the PhD in collaboration with Giovanni Moretti and Sara Tonelli and described in Appendix A, extracts and visualises movements of Shoah victims starting from a dataset modeled using this ontology¹² [Sprugnoli et al., IN PRESS].

The aim of the **Biography Light Ontology** is two-fold: i) encode life events following the 4W model, thus answering questions about what, where, when, who; ii) improve the interoperability among existing standards such as LODE and Bio [Ramos, 2009]. Biography Light introduces the main class `BioEvent` with four subclasses that represent changes in the health of the subject of the biography, his/her relations with other people, changes in location such as migrations, and inventions or discoveries made by the subject. Event properties are borrowed from LODE (e.g., `atPlace`) and from the Event Ontology (e.g., `isAgentIn`).

Bio CRM is a domain specific extension of CIDOC CRM: it provides a general model for representing biographical datasets that can be extended to meet the requirements of specific projects [Tuominen, 2016]. This ontology makes a clear distinction between unary roles of actors, binary relations between actors and events in which actors participate having different roles. Events are described in terms of time, location, participants and other involved resources; moreover, they are organized in an hierarchy distinguishing, for example, ecclesiastical from educational events. Each event type has a corresponding class of permitted roles: Figure 3.3 shows part of the hierarchy of ecclesiastical events together with their roles.

Biography.owl is a lightweight ontology designed to represent biographical facts [Krieger and Declerck, 2015]: its main feature is the tripartite structure with which entities are modeled. More specifically, the most general class `Entity` has three subclasses, that is `Abstract` (describ-

¹²<http://dh.fbk.eu/technologies/lod-navigator>

3.2. PROJECTS

```
bioc:Event  
  :Ecclesiastical_Event  
    :Baptism  
      Roles: :Officiant, :Baptismal_Candidate, :Godfather, :Godmother, :Religion  
    :Confirmation  
      Roles: :Officiant, :Confirmation_Candidate, :Religion
```

Figure 3.3: Two subclasses of the class `:Ecclesiastical_Event` in Bio CRM with their corresponding allowed roles. Example taken from [Tuominen, 2016].

ing concepts and roles), `Object` (describing physical things) and `Happening`. The latter includes both situations and events, the first being static and atomic, the second dynamic and decomposable. Happenings have properties related to their starting and ending date, the agents involved in them, and their location. Particular attention is devoted to pre- and post-conditions of a happening thanks to properties encoding causes and effects. Biography.owl is one of the 18 sub-ontologies composing TMO, an integrated ontology developed within the European TrendMiner project [Krieger and Declerck, 2014].

3.2 Projects

In this Section we provide an overview of projects run in the field of Digital Humanities where the definition, detection and processing of events is central. These projects are initiatives grounded in the collaboration between researchers in computer science, computational linguists, and humanities scholars: in the following we describe the methodologies adopted within each project to deal with events in the historical and cultural heritage domains.

3.2. PROJECTS

3.2.1 FDR/Pearl Harbor

The aim of the FDR/Pearl Harbor project, funded by U.S. National Science Foundation, was to help historians of the Second World War to search and retrieve information from documents (e.g. government correspondence and memoranda) written before the Pearl Harbor attack on 1941 [Ide and Woolner, 2007]. First, about 1,500 documents were digitized and annotated at different linguistic levels: part of speech, verb and noun chunks, named entities, and verbal events. In the second phase, ontological relations between annotated data were derived semi-automatically and then represented using RDF Schemas and OWL. As for event recognition, Ide and Woolner [2004] identified three main types of events, i.e. historical, communicative, and conjectured, but focused only on communicative events. In order to identify this type of event, they extracted all verbs from the corpus, grouped them on the basis of WordNet 2.0 synsets [Fellbaum, 1998] and then assigned one or more FrameNet frame to each group [Baker et al., 1998]. Finally, they selected the group associated with the Communication frame (including its sub-frames). Additional information missing in FrameNet was added manually, for example to distinguish between lexical units with negative or positive valency, such as *condemn* and *acclaim* that are both lexical items of the “Judgment communication” frame. Finally, semantic roles were assigned following a naive representation scheme according to which the pronoun or the named entity of type PERSON occurring before the verb was taken as communicator and the rest of the sentence as topic. No information was instead provided about how the extraction of the addressee was performed. Figure 3.4 shows an example of event recognition and role assignment for the sentence *Mr. Kurusu asked whether this was our reply to their proposal for a modus vivendi*. No evaluation of event recognition is mentioned in the paper.

3.2. PROJECTS

```
COMMUNICATOR: Mr. Kurusu
asked [ask : QUESTIONING: COMMUNICATION]
TOPIC: whether this was our reply to their proposal for a
       modus vivendi.
ADDRESSEE: Secretary Hull
```

Figure 3.4: Event recognition and role assignment in the FDR/Pearl Harbor project. Figure taken from [Ide and Woolner, 2004].

3.2.2 Semantics of History

The Semantics of History project was funded by the Interfaculty research institute CAMERA in Amsterdam with the aim of developing a historical ontology and a lexicon to be applied to a new information system for historical archives. After an analysis of how historical events are realized in different types of texts [Cybulska and Vossen, 2010], the Simple Event Model (see 3.1.1) was identified as the most appropriate model to express the key event dimensions. At a later stage, two tasks were carried out: i) the actual extraction of events, together with their actors, locations and dates, from texts; ii) the selection of events having a historical value. Cybulska and Vossen [2011] created a corpus of 78 Dutch texts about the Srebrenica massacre that happened in 1995 and processed it using the KYOTO framework 7 [Vossen et al., 2008a]. This framework includes a pipeline of NLP modules and allows for the development of so-called Kybot profiles, that is XML files that specify patterns to be used to detect information of interest. More specifically, all the documents were automatically annotated with part of speech information and lemmas, parsed with a dependency parser, and tagged with semantic labels taken from ontological classes and the Dutch WordNet [Vossen et al., 2008b]. In addition, a named entity recognition system was used to identify dates and geographical names. In order to define semantic classes relevant for the identification of historical events, the authors manually extracted events, actors, locations and

3.2. PROJECTS

dates from 5 documents and associated them with the corresponding WordNet synset. The mapping between semantic classes and WordNet synsets, combined with morphosyntactic information, was used to develop Kybot profiles to detect events, actors, locations and dates. As for events, only conflict-related and motion actions were considered historically relevant. In particular, both verbal (e.g. *to deport*) and nominal (e.g. *genocide*) actions, as well as actions with a syntactic object (e.g. *start the aggression*) were extracted with the Kybot profiles. All the other events, such as cognitive events, were considered unimportant from the historical point of view. Evaluation was performed on 5 documents manually tagged by two annotators with a very high inter-annotator agreement (0.91 Kappa). The system achieved an overall precision of 57%, a recall of 49%, and an F-measure of 0.53.

3.2.3 Agora

The Agora project supported by the Center for Advanced Media Research of Amsterdam, aimed to enrich metadata in museum collections through the automatic extraction of historical event names from unstructured Dutch texts [Van Den Akker et al., 2010]. To this end, historical events were extracted using standard information extraction techniques and modeled adopting SEM. Segers et al. [2011] selected 3,724 Wikipedia articles on historical topics and performed a three-step procedure. In the first step, actors (i.e. persons and organizations) and locations were extracted from the corpus using the Stanford Named Entity Recognition system [Finkel et al., 2005] adapted for Dutch while dates were identified by means of a set of regular expressions. In the second step, a list of frequent patterns (e.g. *after the, during the*) occurring with one hundred seed events (e.g. *French Revolution*) was collected from the Web and applied to the corpus. After filtering out the event candidates on the basis of a threshold,

3.2. PROJECTS

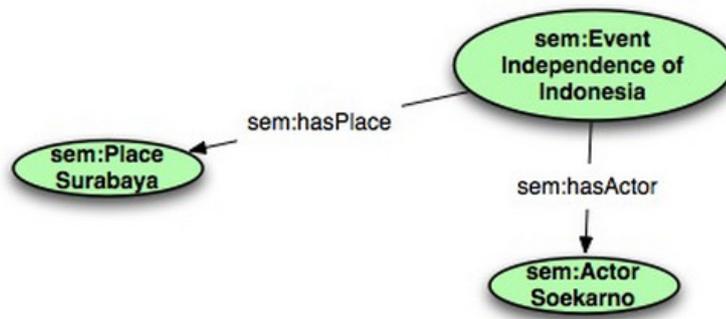


Figure 3.5: SEM instance filled with information: the event name *Independence of Indonesia* is associated with an actor *Soekarno* and a place *Surabaya*. Image taken from [Van Erp Marieke and Schreiber., 2011].

2,444 unique events were identified with a precision of 56.3%. In the last step, events extracted with the pattern-based method in the previous phase were associated with actors, locations and dates by checking the occurrence of each combination (i.e. event name - actor; event name - location; event name - date) on the Web and assigning a score to them. A manual evaluation of the resulting associations was performed: the precision was 71.9% for event names, 45.6% for actors, 51.5% for locations, and 51.5% for dates. Events and entities extracted from the Wikipedia corpus were used to populate a set of SEM instances (see Figure 3.5) forming an historical event thesaurus. Such thesaurus was included in a browser created to access the collections of the Rijksmuseum Amsterdam and the Netherlands Institute for Sound and Vision facilitating event-driven browsing and search [Van Erp Marieke and Schreiber., 2011].

3.2.4 DIVE+

DIVE+, the follow-up project to Agora, has a similar aim (i.e., support event-driven browsing and search within heterogeneous collections) and general approach (i.e., combination of Semantic Web technologies and NLP approaches to improve findability and digital hermeneutics in cul-

3.2. PROJECTS

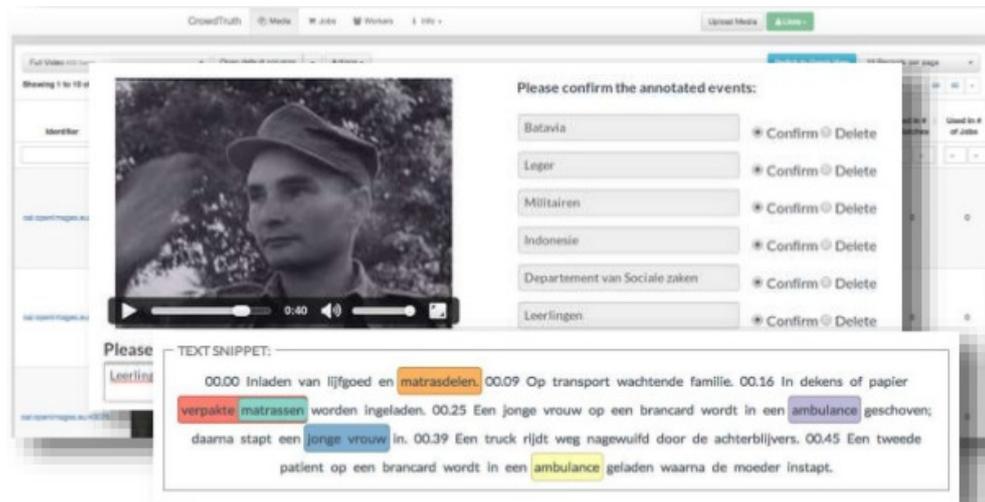


Figure 3.6: Validation and annotation of event identification in texts and video in the DIVE+ project.

tural heritage collections) but adds extensive use of crowdsourcing [De Boer et al., 2015]. The project deals with three types of content, each associated with its metadata record, coming from different Dutch institutions: video of news broadcast, radio news scripts, images of cultural heritage objects. Textual descriptions and descriptive metadata of this content are processed using named entities recognition tools, such the one developed in the OpeNER project¹³ to automatically extract events, locations, persons and concepts. In addition to named events (e.g. *Second South-New Guinea Expedition*), others are taken from structured metadata schema that are based on ontologies modeling events as those described in previous Section. The CrowdTruth platform [Inel et al., 2014] is then used to refine the output of automatic tools and to perform human annotation and enrichment (see Figure 3.6¹⁴. In particular, this manual enrichment phase helps to add missing events and missing links between events and participating entities. The combination between tools and crowdsourcing led to the identification

¹³<http://www.opener-project.eu/>

¹⁴Image taken from the presentation “DIVE+ and Events” at the EVENTS2017 workshop: <https://www.slideshare.net/vdeboer/dive-and-events-at-events2017>

3.2. PROJECTS

of more than 199,000 events and more than 685,000 links between events and entities (media objects, people, places, concepts). These results are consolidated to RDF and events are modeled adopting the Simple Event Model. The data, forming a large knowledge graphs, is stored in a public triple store and the resulting linked data cloud is visualized through a web based portal¹⁵.

3.2.5 NanoHistory

On May 1, 1658 Chad told Abe he saw Ben kill Dean in Leicester Square

```
[May 1, 1658]->[Chad->spokeTo->Abe]->content ->[[Chad->witnessed->[Ben->killed->Dean]]->in->Leicester Square]
```

- 1: Ben killed Dean
- 2: Chad witnessed {1}
- 3: {1} in Leicester Square
- 4: Chad spokeTo Abe
- 5: {4} content {3}
- 6: {5} hasDate May 1, 1658

Figure 3.7: Manual modeling of events in the Nanohistory project.

NanoHistory is an ongoing project aiming at developing network representations of the past that document the connections between different kinds of entities, such as people, organizations, places, and physical objects. The notion of event is central in the project but the provided definition is very vague: “Defining an event can be messy, but NanoHistory sees it as something that doesn’t necessarily have a title or a name - it just is.”¹⁶. The focus is on events encoding interactions and relationships between an agent and an object. In the current phase of the project, the modeling is done manually following a nano-level granularity: descriptions of interactions are broken in the smallest possible units of text following the syntactic structure of the sentence, each unit representing an event. Events are identified on the basis of a controlled vocabulary of verbs that currently consists

¹⁵<http://dive.beeldengeluid.nl/>

¹⁶<http://www.nanohistory.org/>

3.2. PROJECTS

of 359 entries grouped in an event typology with 11 classes: Art, Components, Conflict, Economic, Education, Familial, Geography, Governance, Legal, Life Cycle and Travel. Events are unnamed to leave historians the task of interpreting them within their personal research investigation. For this reason, labelled historical phenomena, called Episodes, are considered as pseudo-entities, not as a sub-type of events. On the contrary, titles (*the King of French*) and occupations (*printer*) are documented as events. Scholars contribute to the project through a collaborative platform linked to several Linked Open Data resources, such as Geonames and Europeana, from which users can harvest external information. Figure 3.7 presents an example of manual event model in NanoHistory taken from the project website¹⁷.

3.2.6 Bringing Lives to Light: Biography in Context

Bringing Lives to Light is a project of the Electronic Cultural Atlas Initiative (ECAI) at the University of California: its goal is to design, develop and evaluate tools that can improve the understanding of biographical texts by connecting life events to contextual information, that is their location, time of occurrence and related archival materials [Buckland and Ramos, 2010]. Different datasets and sources of information are taken into account: the digital texts provided in the online Biographical Directory of the United States Congress¹⁸, the manually compiled chronology of Emma Goldman itinerary, the scanned page image of Irish texts. Event identification is carried out following two approaches: i) manual modeling on the basis of the Biography Light Ontology (see Section 3.1.2); ii) automatic extraction of named entities using the GoldeGATE editor [Sautter et al., 2007] and

¹⁷<https://www.nanohistory.org/about/networked-events/>

¹⁸<http://bioguide.congress.gov/>

3.2. PROJECTS

linking to authoritative identifiers, such as GeoNames¹⁹ for locations [Gey et al., 2008]. The result of the first approach is the extraction of event factoids modeled as instances of biographical event classes; an example is given below.

Text: Robert George Collier Proctor (1868-1903), bibliographer, was born in Budleigh Salterton, Devon, on 13 May 1868. He was educated at a preparatory school in Reading and at Marlborough College, before joining Bath College in 1881.

Event factoids:

- **ChangeOfHealth:**
 - birth, 1868-05-13, Budleigh Salterton, Devon
- **ChangeOfSocialRelation:**
 - studied at Marlborough College, before 1881
 - studied at Bath College in 1881

Furthermore, the automatic extraction and enrichment method led to the development of websites, such as the one exemplified in Figure 3.8, where people’s events are represented in a geo-temporal browser. No evaluation of the accuracy of the automatic processing is reported in the project publications.

¹⁹<http://www.geonames.org/>

3.2. PROJECTS

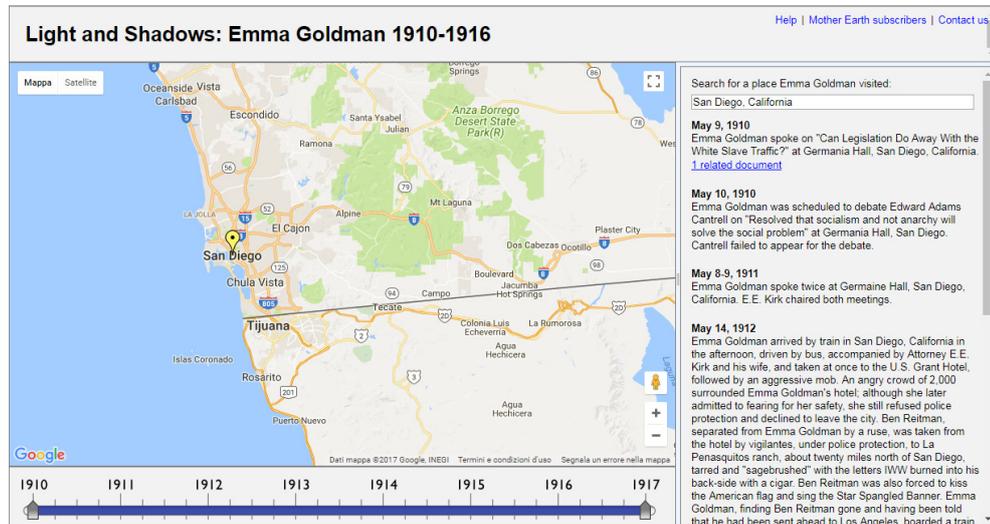


Figure 3.8: Geo-visualization of the chronology of events in Emma Goldman’s life. Screenshot from <http://metadata.berkeley.edu/emma/>.

3.2.7 BiographyNet

Sponsored by the Netherlands eScience Center, the BiographyNet project²⁰ aims at improving the way historians use biographical texts for their research through the adoption of computational methods. The project uses data of the Biographical Portal of the Netherlands²¹ which contains short biographies of 76,000 individuals. A semantic knowledge base of relations between people, places and events is built by extracting information with NLP techniques, converting this information and metadata to RDF and linking all to external resources [Fokkens et al., 2014]. The automatic data processing includes the identification of named entities of people, locations and organizations, the detection of temporal expressions, the disambiguation of word senses, the extraction of predicates encoding events connected to their frames and the labelling of roles for each event participant. The output of the data processing phase is converted into RDF triples adopting

²⁰<http://www.biographynet.nl/>

²¹<http://www.biografischportaal.nl/>

3.2. PROJECTS

a schema compatible with the Europeana Data Model (see Section 3.1.1). This schema poses particular attention to the provenance of each extracted information by using the PROV Data Model (PROV-DM). PROV-DM allows the specification of the information extracted from the metadata or from automatic processing: in the latter case, the information is linked to the corresponding tokens in the original source and the overall performance of the automatic system is indicated, thus historians can verify the level of trustworthiness of the data they are analyzing [Ockeloen et al., 2013, Fokkens et al., 2018].

3.2.8 Semantic Biographies Based on Linked Data

The Semantic Computing Research Group of Aalto University (Finland) is carrying out a set of experiments and projects on the linking, enrichment and visualization of biographies. Similarly to the project Bringing Lives to Light presented in Section 3.2.6, this initiative wants to improve the reading experience of biographies by providing the users with a rich reading context, in other words by adding external content to biographical datasets. A first experiment, called National Semantic Biography of Finland, takes as input data the short biographies published in the Finnish National Biography²² and works on a single type of event. An event extractor is used to identify snippets of texts containing words expressing creation events, dates written in numbers, named entities of type location and a reference to the name of the subject person of the biography. Extracted information is then transformed in RDF following the Bio CRM model (see Section 3.1.2) and linked to several external resources such as GeoNames and Wikipedia [Hyvönen et al., 2014]. This approach has been applied also to the digitized historical register of the Finnish high school “Norssi” that includes information about the studying lives of more than 10,000 alumni

²²<https://kansallisbiografia.fi/english>

3.2. PROJECTS

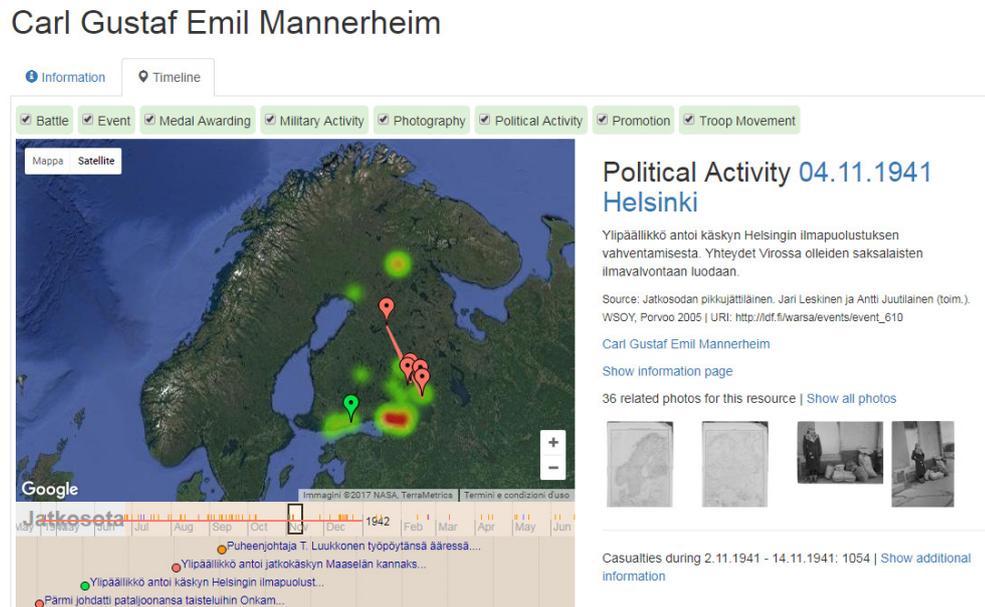


Figure 3.9: Event visualization in the Person perspective of WarSampo portal at <https://www.sotasampo.fi/en/>.

[Hyvönen et al., 2017]. However, in this case, the precise structure of each biography allowed to use simple regular expressions to extract information. Finally, the WarSampo project proposes different perspectives to the history of World War 2 and events play a crucial role to reassemble the narratives of battles, troop movements and soldiers' lives. In this project, events are defined as “the semantic glue for data linking” [Hyvönen et al.]: they are modeled following the CIDOC CRM model (see Section 3.1.1) and annotated using a specific domain ontology of war time incidents [Hyvönen et al., 2015]. This annotation is carried on using the ARPA tool, a web service that automatically links text descriptions to a vocabulary or an ontology [Mäkelä, 2014]. The outcome of the project is a web portal in which the user can browse the content of several datasets on the basis of different but interconnected perspectives. For example, Figure 3.9 shows a person's page: events of the life of the Field Marshal Carl Gustaf Emil Mannerheim are categorized in classes (e.g. military activity, political activity) that can

3.3. WEAK POINTS OF CURRENT EFFORTS IN DH

be selected or not so to make them appear/disappear in the timeline and their location display/hide in the map.

3.3 Weak Points of Current Efforts in DH

By analyzing the projects described in the previous Section, and whose main approaches are summarized in Table 3.3, we observe that, unlike what happened in other domains such as the clinical one, no real 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. Some projects only take into consideration verbal events, as in the case of the FDR/Pearl Harbor project and NanoHistory, or prefer to identify events manually, such as in Bringing Lives to Light and NanoHistory. On the other side, in Agora and DIVE+ 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 narrowing the extraction of events to a limited set of types. For example, the FDR/Pearl Harbor project focused only on communication events, the National Semantic Biography of Finland project on creation events and Semantics of History only on conflict-related and motion actions. Another questionable point concerns the idea of a priori automatic selection of events on the basis of their historical importance as in the Semantics of History project. Indeed, in historical investigation, the distinction between an important event and one with no value is never definitive and depends on the research questions and on the sources to be analyzed [Marrou, 1954b].

The scarcity of corpora fully annotated with temporal information is

3.3. WEAK POINTS OF CURRENT EFFORTS IN DH

an additional weak point of current NLP research for historical texts that has a strong negative impact on the development of NLP systems, given that no annotated corpora means no training and test data. Indeed, files tagged within the projects described above have not been publicly released. Two notable exceptions are the ModeS TimeBank [Nieto et al., 2011], containing Spanish texts from the 18th 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 20th century already described in Section 2.2.4. Both were manually annotated following a language-specific adaptation of TimeML but ModeS TimeBank was employed only for theoretical studies on the evolution of the Spanish language, while the De Gasperi corpus was used to measure the performance of event extraction

	ONLY VERBS	MANUAL ANNOTATION	ONLY NAMED ENTITIES	LIMITED TYPES	CROWDSOURCING	FRAME APPROACH	LINKING
FDR/Pearl Harbor	X			X		X	
Semantics of History				X			
Agora			X				
DIVE+			X		X		X
NanoHistory	X	X		X			X
Bringing Lives to Light	X		X				X
BiographyNet						X	X
National Semantic Biography of Finland				X			X
Norssi Register				X			X
WarSampo				X			X

Table 3.1: Summary of the main characteristics of DH projects described in Section 3.2.

3.4. CHAPTER SUMMARY

systems on historical texts within the EVENTI evaluation exercise (see Section 2.2.4 for more details on EVENTI).

3.4 Chapter Summary

In this Chapter we present a survey and an analysis of the state of the art in event definition and processing in the Digital Humanities. In this field, a lot of effort is devoted to the modeling of events through the use of Semantic Web ontologies: for this reason, we describe some of these ontologies, highlighting how events are represented through their properties and relations in history and cultural heritage. We also provide an overview of the schemata developed to represent life events in biographies, taken as a notable example of data much investigated by historians being unique sources of historical knowledge. Several projects are then described presenting the results of interdisciplinary collaborations: by analyzing the methodologies adopted by these projects, we propose a discussion about the main limitations of recent and current approaches to event detection and processing in DH. In particular, we show that a lot of work has been done to develop ontologies for dealing with humanistic texts whereas, on the contrary, NLP techniques have not been fully exploited yet and that the historians' perspective should be taken into account to make the output of NLP tools meet their needs.

In the next Chapter we will focus on this last aspect and we will illustrate the questionnaire we conducted to help define what counts as an event for scholars in the historical domain. We will also report and analyse the obtained results in order to understand whether historians' definition is compatible with the event definitions and the approaches adopted in the NLP community.

3.4. CHAPTER SUMMARY

Chapter 4

Survey on Event Definition and Annotation

In this Chapter we present a case study we conducted taking the perspective of history scholars, i.e. researchers from the Humanities that typically deal with events in their daily activity. In particular, we try to address the following questions:

- (i) was all the work devoted to event processing with IE techniques useful to serve real historical investigation?
- (ii) were the various definitions of events provided over the years compatible with research practices adopted in other communities?
- (iii) 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 and annotation’ exercise. The outcome of this study, published in [Sprugnoli and Tonelli, 2017], highlights the difficulties in shifting from a linguistic-driven perspective to the historical one, where a more abstract definition of events is prevalent.

4.1. WHAT IS AN EVENT IN HISTORY?

4.1 What is an Event in History?

According to the American Managing Editor and Professor Charles Angoff, “History is a symphony of echoes heard and unheard. It is a poem with events as verses.” This evocative definition highlights the complex nature of History, made of traces to be retrieved and interpreted, and the central role that events play in the understanding of the past. Anyway, as shown in the previous Chapter, past projects trying to apply NLP techniques to historical investigation have adopted heterogeneous approaches, and there has been no real effort among history scholars to standardize event definition taking into account 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.

Questionnaires have been already employed in NLP to carry out user requirements studies or to discover trends in the use of a specific technique, [Allen and Choukri, 2000, FLaReNet Working Group, 2010, Oostdijk and Boves, 2006, Tomanek and Olsson, 2009]. European research infrastructure consortia such CLARIN-ERIC and DARIAH have proposed web surveys at national and international levels as well, for examples to identify digital scholarly practices in the Arts and Humanities and the needs of ancient Greek scholars in Italy [Costis et al., 2017, Monachini et al., 2017]. However, to the best of our knowledge, our questionnaire is the first one on this topic, whose outcome can potentially enrich the current theoretical discussion on the nature of events. Besides, it was seen as a preliminary step towards the definition of annotation guidelines for developing NLP

4.1. WHAT IS AN EVENT IN HISTORY?

tools in this domain (see Chapters 5 and 6).

4.1.1 Questionnaire Description

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 18 questions, among which were both closed questions (allowing for a statistical summarization of responses) and open questions (giving participants the possibility to elaborate upon answers to closed questions and to leave feedback). Only the examples taken from historical documents were different but with the same range of linguistic phenomena to investigate. Questions, in both languages, are reported in Appendix B.

The questionnaire was distributed via social media (i.e. Twitter and LinkedIn), mailing-lists (e.g. the Humanist Discussion Group¹) and targeted emails to individual historians, professional associations (e.g., the Australian Historical Association²) and research centers (e.g., Institute of Historical Research at the University of London³).

The questionnaire consisted of three main parts. The first part aimed at shedding light on the way historians interpret the notion of event. The second part of the questionnaire focused on assessing the interest of historians towards the use of Natural Language Processing tools in support of historical investigations. In the third part we collected demographic information (e.g. age and nationality). 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.

¹<http://dhumanist.org/>

²<http://www.theaha.org.au/>

³<http://www.history.ac.uk/>

4.1. WHAT IS AN EVENT IN HISTORY?

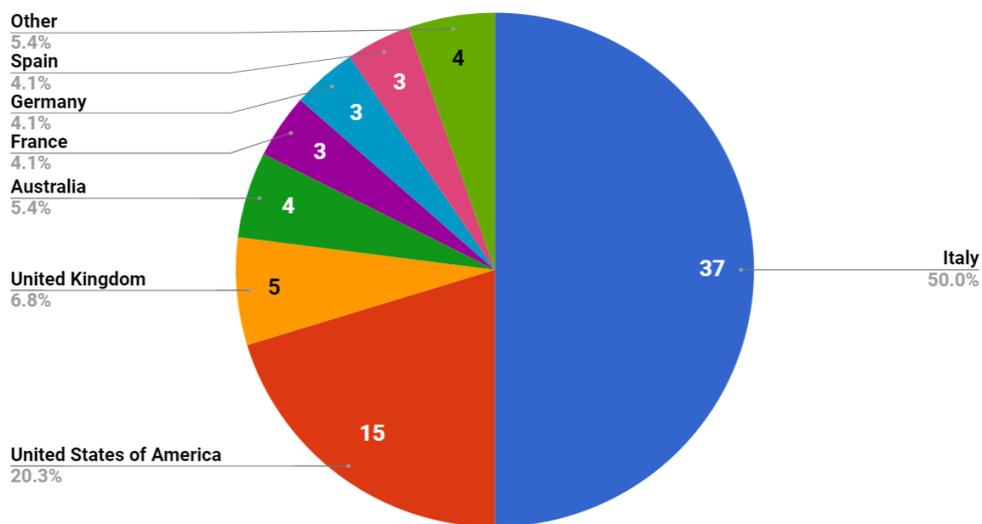


Figure 4.1: Answers to the question “In which country do you live?”. Belgium, Canada, Greece and Republic of Serbia are the countries gathered in the *Other* category.

4.1.2 Questionnaire Results

After two months from its launch, 74 historians participated in the survey with a balanced distribution across different age groups (from 20 to above 60 years old). As displayed in Figure 4.1, we had a strong feedback from Italy (50%) but also other countries were well-represented, in particular the US (21.7%), the UK (7.1%) and Australia (5.7%). Figure 4.2 shows instead that historians from more than 10 research fields took part in the questionnaire with a prevalence in social and cultural history (30.1%), political history (15.4%) and intellectual history (11.5%) thus providing insight into current research practices in the domain on interest.

As for the attitude towards NLP, results showed a generally positive opinion about this type of research and its methods: although only 6.8% of participants said they used NLP in their research, the majority of the respondents were interested in knowing more about it and only 3 out of 74 respondents declared a complete lack of interest in NLP. Respondents also stated to already know several NLP methods and tasks: lemmatisation,

4.1. WHAT IS AN EVENT IN HISTORY?

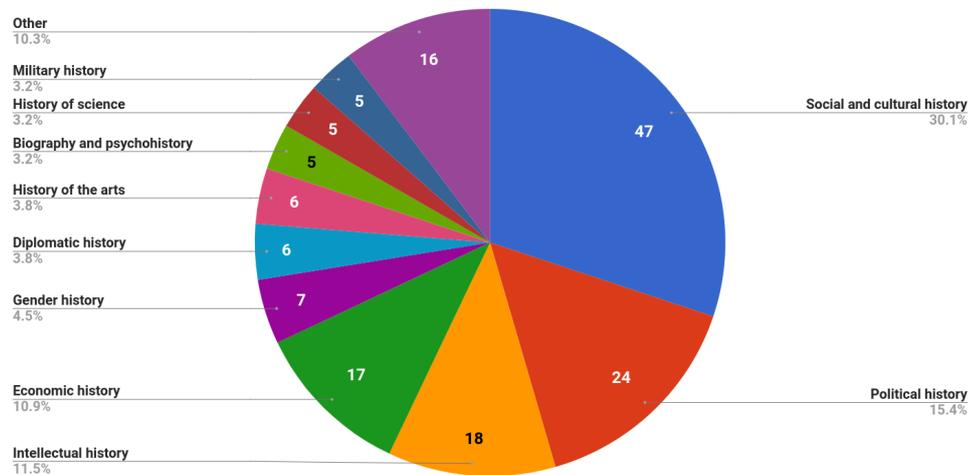


Figure 4.2: Answers to the question “How would you define your field of research?”. Respondents could give more than one preference. Under the *Other* category, 11 additional fields were indicated: i.e., digital humanities, archaeology, history of books, history of religions, epistemology of history, spatial history, legal history, methodology of the historical research, history of Christianity, history of publishing, history of the East.

PoS tagging, Named Entity Recognition, tokenization and parsing were the most popular answers, followed by sentiment analysis, topic modeling and text simplification. Anyway, answers were cautious when coming to the possibility of actually adopting an NLP tool to identify temporal information within historical texts in everyday research practice: 38.6% of the respondents were not sure, 10% were not interested in these tools, and the others were willing to try but only with appropriate training.

Three questions were then used to shed light on the notion of “event” for historians. 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.

The sentences were different in the English and in the Italian questionnaire but they contained the same linguistic phenomena: negated verbs

4.1. WHAT IS AN EVENT IN HISTORY?

SENTENCES		ST	MT	Most Common Extents
<i>Today, once again, the independence of the Western Hemisphere is menaced from abroad</i>	V	17%	8%	today menaced
	NV	51%	24%	independence
<i>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</i>	V	4%	13%	conference end of the Korean war
	NV	28%	55%	disarmament
<i>I think we can work that out with the advice of the Ways and Means Committee</i>	V	20%	28%	advice work that out
	NV	36%	16%	think

Table 4.1: 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.

(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*). Questionnaire participants could annotate single words or expressions conveying events, but also provide no annotation. We report in Table 4.1.2 the English sentences, taken from J. F. Kennedy’s public speeches⁴ while Table 4.1.2 shows Italian sentences extracted from public documents by Alcide De Gasperi. The percentages listed in the tables 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

⁴Available at http://www.presidency.ucsb.edu/1960_election.php.

4.1. WHAT IS AN EVENT IN HISTORY?

SENTENCES		ST	MT	Most Common Extents
<i>Man mano che avanzavano, i soldati andavano a prender posto nei vagoni del lunghissimo convoglio</i>	V	36%	29%	andavano soldati
	NV	29%	6%	andavano a prender posto
<i>Invece la guerra non è ancora finita e la pace sembra ancora lontana</i>	V	12%	26%	guerra pace
	NV	29%	6%	la guerra non è ancora finita
<i>Ella voglia la prego aggiungere che ci rendiamo perfettamente conto della delicatezza della questione che incide sui rapporti stessi fra i grandi alleati</i>	V	17%	28%	questione rapporti
	NV	48%	7%	ci rendiamo conto

Table 4.2: Italian 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.

contained in the expression. In particular, all expressions containing at least one verb were considered verbal.

As for English, in the first and the third sentence a high percentage of respondents (37% and 68% respectively) did not detect any event, probably because these sentences contain a state (i.e. *independence*) and an opinion (i.e. *I think...*). In the second sentence only the 3% of respondents did not annotate any event, probably because it includes a named event (*Korean war*). Percentages are high also for Italian: 33% in the first sentence, 47% in the second and 51% in the third.

In English, the majority of the identified events are non-verbal (e.g., *today, independence, conference, end of the Korean war, advice*): 75% in

4.1. WHAT IS AN EVENT IN HISTORY?

the first sentence, 83% in the second, and 52% in the third. This contrasts with the outcome of the experiment reported by Hatzivassiloglou and Filatova [Filatova and Hatzivassiloglou, 2003], in which nouns such as *war* and *earthquake* were never identified as events by a group of students annotating news. Non-verbal events are the majority also in the second and third Italian sentences (62% and 55% respectively) while in the first sentence verbs (e.g., *andavano*, *avanzavano*) prevail having a percentage of 65%. In the first and third Italian sentence, several nouns and adjectives are marked as events though they identify objects (*vagoni*) and participants (*alleati*) or adjectives (*lunghissimo*, *grandi*) that characterize them⁵.

Events consisting of more than one token are annotated frequently. As for English, 32% of annotated events are multi-token in the first sentence, 68% in the second, and 44% in the third. As for Italian, multi-token expressions are annotated in the 35% of the cases in the first sentence, 27% in the second and 36% in the third. Some of these multi-token extents correspond to entire clauses, e.g., *This country has not been prepared* and *i soldati andavano a prender posto nei vagoni del lunghissimo convoglio*. This high number of multi-token events goes against the TimeML and RED minimal chunk rule for tag extent, according to which only single tokens are to be annotated as events.⁶ 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. More-

⁵Please note that the percentages related to Italian sentences do not take into consideration few cases in which respondents have indicated as events words not present in the sentences but derived from a re-elaboration: e.g., *treno* and *battaglia* in the first sentence and *diplomazia* in the third.

⁶The 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 75 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., 2011a].

4.1. WHAT IS AN EVENT IN HISTORY?

over, ACE, ERE and Event Nugget allow the annotation of multi-token event triggers (the latter also in discontinuous cases).

Point 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. Among the considered standards, the multi-token annotation of continuous and discontinuous 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 labeled 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.⁷ The ratings included 4 possible values, i.e. “very important”, “somewhat important”, “not important”, and “don't know”. Figure 4.3 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

⁷<http://rbedrosian.com/event.htm>

4.1. WHAT IS AN EVENT IN HISTORY?

Nugget consists 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 2.2). 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 and emotion. The USAS [Rayson et al., 2004] and the Historical Thesaurus of English tagsets [Kay et al., 2009b] contain 21 and 37 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, while it is more relevant from a linguistic perspective. In fact, TimeML encodes this property through subordination links (SLINKs) 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 (TLINKs), corresponding to 13 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-2015.⁸ In TimeML, the TLINK 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 degree to which an event can be pinpointed to a particular time or period, and the *Frequency* property. In the MUC Scenario Template as well as in ACE, ERE and Event Nugget, temporal relations between events or

⁸<http://alt.qcri.org/semeval2015/task4/>

4.1. WHAT IS AN EVENT IN HISTORY?

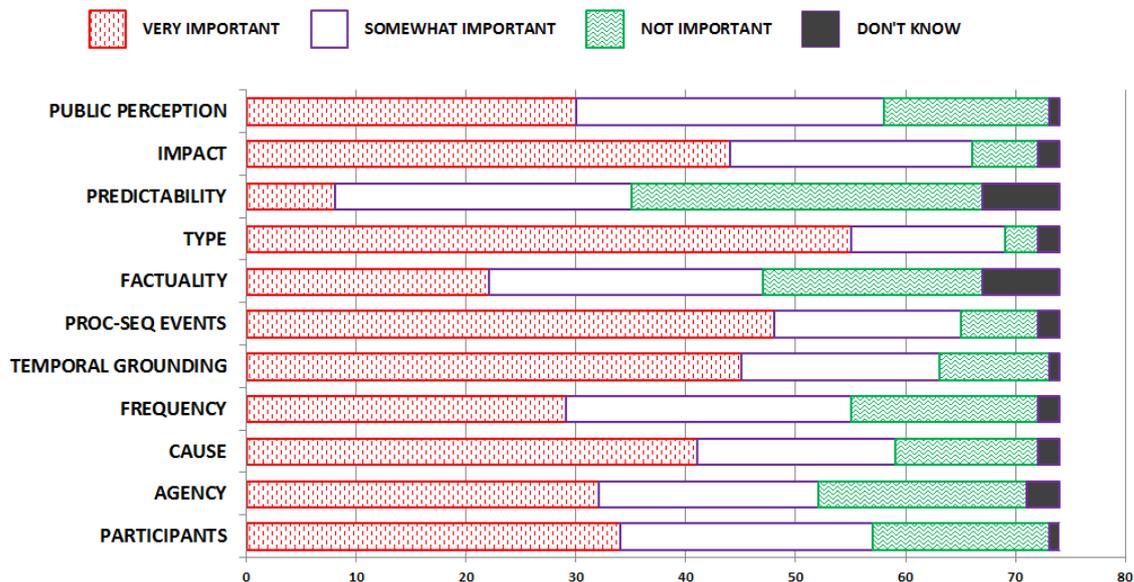


Figure 4.3: What are the most important properties for a historian in order to understand if a word (or a set of words) expresses a relevant event.

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 2.2). 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 TLINK (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, Mirza et al., 2014, Mirza and Tonelli, 2016]. 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 an-

4.1. WHAT IS AN EVENT IN HISTORY?

notation, but they are currently not included in the guidelines. However, on the contrary, causal relations play an important role in the RED guidelines [Ikuta et al., 2014]. 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 et al., 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 2.2). Research on semantic roles can provide much guidance in this respect, for example by taking inspiration from PropBank [Palmer et al., 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.

Point 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 points. The English questionnaire presented the following passage taken from a speech uttered by J.F.Kennedy during the 1960 presidential campaign, while the Italian questionnaire had a sentence from a news

4.1. WHAT IS AN EVENT IN HISTORY?

report written by Alcide De Gasperi in 1914:

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]₁ up]₂ the next day. Our Ambassador did not [[show]₁ up]₂ for 9 months.

A Londra si [caricano]₁₋₂ forse le tinte per [convincere]₁₋₂ la Germania che la Russia [vuol]₁₋₂ proprio [[fare]₁ sul serio]₂; a Berlino [incomincia]₁₋₂ a [scemare]₁₋₂ l'ottimismo degli altri giorni, da Vienna non si è [comunicato]₁₋₂ alla stampa nessuna notizia.

In the annotation marked with [...] ₁ only single tokens are annotated as events following the TimeML specifications. Moreover the states *independence* and *ottimismo* are not annotated. The option marked with [...] ₂ proposes looser criteria, annotating both multi-token event expressions and states. By merging the answers given in the two versions of the questionnaire, it emerged that only 5% of participants preferred the first annotation, 61% 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

4.1. WHAT IS AN EVENT IN HISTORY?

annotation rules, as in TimeML, would be extremely critical because such rules could not cover all the information needs of historians.

Point 3. Point 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.

4.1.3 Discussion

On the basis of historians' replies to our questionnaire and of the analysis of the state of the art both in the NLP and in the DH domain, we can now answer the questions posed at the beginning of this Chapter:

(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.

(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 in-

4.2. CHAPTER SUMMARY

formation objects characterized by many properties. These can be cast by combining different NLP analyses providing rich semantic information, such as semantic role labeling, 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 to 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 analyzing [Marrou, 1954a].

4.2 Chapter Summary

In this Chapter we present an online questionnaire we have designed and distributed in order to better understand historians' notion of events and their requirements. More specifically, we describe the structure of the questionnaire and its results. Its outcome shows that a careful adaptation of existing annotation schemes is necessary to meet the requirements of experts in the historical domain.

In the next Chapter the findings of the questionnaire will be used to

4.2. CHAPTER SUMMARY

develop new event annotation guidelines and a corpus of historical texts will be annotated and publicly released.

Chapter 5

Events in Historical Texts: Guidelines and Manual Corpus Annotation

On the basis of the outcome of the questionnaire described in the previous Chapter, we have developed new annotation guidelines focused on the identification and classification of textual mentions denoting all types of (punctual or durative) actions, processes and states. The goal is to meet the requirements of history scholars emerged from the questionnaire in terms of event definition, extension and grammatical realization so as to effectively apply NLP and linguistic annotation to their research area. Moreover, we focus on event classification since the type resulted as the most important property of an event.

As reported in Section 5.1, in our guidelines we have adopted a wide definition of event by referring to the Bach’s notion of eventuality¹ [Bach, 2008] and, from the linguistic point of view, by taking into consideration different parts of speech and syntactic constructions. In addition, we have defined a set of 22 semantic classes with the aim of providing an exhaustive categorization of events, thus overcoming the limited classifications

¹In this Chapter we use the terms “event” and “eventuality” interchangeably.

5.1. EVENT ANNOTATION GUIDELINES

proposed by other initiatives such as ACE and Rich ERE.

We followed the guidelines thus defined to manually annotate a corpus of historical news and travel writings freely available for research purposes. This resource, called *Histo Corpus*, is described in Section 5.2: details on the inter-annotator agreement are provided together with an analysis of the final annotated dataset also by looking at the correspondence between the number and type of events occurring in a text and the semantic and functional characteristics of the text itself. This idea is based on the notion of Content Types that we have developed in collaboration with colleagues from Trento and The Netherlands. To this end, we have added another annotation level to the *Histo Corpus* by tagging Content Types at clause level.

Our work on Content Types has been published in the Proceedings of EACL 2017 [Sprugnoli et al., 2017a] while two studies on historical travel writings, a genre included in the *Histo Corpus*, are published in [Sprugnoli et al., 2017c] and [Sprugnoli, 2018].

5.1 Event Annotation Guidelines

In this Section we detail the annotation guidelines designed to detect and classify event mentions in texts.

5.1.1 Event Linguistic Realization

Syntactically, the linguistic elements which may realize an event are the following:

- verbs in both finite and non-finite form;

(1) *she expected to be attacked*

5.1. EVENT ANNOTATION GUIDELINES

- past-participles in the nominal pre-modifier position that represent resultatives events. Interpreted as a state, the following example can be paraphrased as “the state of having been imprisoned”;

(2) *an **imprisoned** criminal*

- present-participles in the nominal pre-modifier position that represent in-progress events. In the following example, the modifier describes an event in progress so that it can be paraphrased as “the audience that is smiling and applauding”;

(3) *a **smiling** and **applauding** audience*

- adjectives in predicative position;

(4) *the museum itself **was damp***

- nouns which can realize eventualities in different ways:

– deverbal nouns denoting an activity or an action;

(5) *the **running** of these ferries*

– nouns which have an eventive meaning in their lexical properties even if they not derive from verbs;

(6) *delegates of Russia against the **war***

– post-copular nouns;

(7) *it **was a lie***

– nouns which normally denote objects but which are assigned an eventive reading either through the process of type-coercion [Pustejovsky, 1991b], or through the processes of logical metonymy and coercion induced by temporal prepositions.

(8) *I am finishing this **letter** rather hurriedly*

5.1. EVENT ANNOTATION GUIDELINES

- pronouns related to previously mentioned events.

Differently from the Rich ERE annotation, we do not annotate implied events indicated by nouns like *murderer* and *protestor* so to make a clear distinction between events and entities and avoid confusion.

The factuality status of events does not impact on the annotation: all events have to be annotated whether they are presented as a fact, a counterfact or a possibility.

(9) *here we **saw** all manner of beautiful and hideous creatures*

(10) *Professor Pais wrote a paper on this which I have not **seen***

(11) *one may **see** small towns*

5.1.2 Event Extent

Eventualities have different extents: the annotation of single-token, multi-token and discontinuous expressions is allowed as detailed below.

Finite and non-finite verb forms: annotate only the verbal head without auxiliaries of any form (multiple, modal, negative).

(12) *you **wish to know**= 2 annotations*

(13) *having been **destroyed** by the father*

(14) *we could **appreciate** to-day*

(15) *I do not **understand** Beloch*

Phrasal verb constructions: the main verb should be annotated together with the particle and/or the preposition forming the phrasal verb because they form a single semantic unit whose meaning cannot be understood by looking at the meaning of each single part. In case the verb and the preposition are separated having the direct object in the middle, as in (17), a discontinuous annotation should be performed.

5.1. EVENT ANNOTATION GUIDELINES

(16) *I abjectly **stepped into** his cab*= 1 annotation

(17) *he would have **carried it off** to France*= 1 annotation

Light verbs: the whole predicate formed by the main verb and the following expression, usually a noun, is to be annotated even if not continuous.

(18) ***make her a visit***= 1 annotation

(19) ***get a snap-shop***= 1 annotation

Copular constructions: the literature [den Dikken and O’Neill, 2016] distinguishes different types of copular constructions on the basis of a taxonomy of four copular elements: (i) support copula; (ii) predicational copula; (iii) equative copula; (iv) silent copula. The first two cases are to be annotated with a multi-token span including both the copula and the whole copula complement (20, 21). As for equatives, whose linguistic status is unclear [Mikkelsen, 2005], only the copula should be annotated (22). No annotation is provided in case of silent copula, given that the linguistic realization of the copula is missing: in (23) only the main verb is annotated.

(20) *Our welcome to Genoa **was not cheerful***= 1 annotation

(21) *Propertius **was a contemporary** of Virgil*= 1 annotation

(22) *Dr. Jekyll **is** Mr. Hyde*

(23) *Petrarch **considered** this tomb (to be) sufficiently important to plant a laurel*

Inverse copular constructions [Moro, 1997] should be annotated as well, taking into consideration the reversed word order from the canonical *subject-copula-predicative expression* to *predicative expression-copula-subject*:

(24) ***an interested onlooker was** former Coroner Gustav Scholer*= 1 annotation

5.1. EVENT ANNOTATION GUIDELINES

Please note that the verb *to be* is not the only English copula to take into account²:

(25) *I felt strange*= 1 annotation

Periphrastic causative constructions: are composed of a causative verb such as *cause*, *get*, *have* or *make* combined with another verb to express causation [Kemmer and Verhagen, 1994]. These two verbs should be annotated separately.

(26) *urging him to make his brother drive more carefully*= 2 annotations

Fixed expressions: phrases, idioms, nominal expressions whose meaning cannot be understood from the individual meanings of their elements have to be annotated as a unique mention.

(27) *in order to get rid of him*= 1 annotation

(28) *I would not have a leg to stand on*= 1 annotation

(29) *a hostile air raid this evening*= 1 annotation

Nouns: can be annotated within a multi-token or discontinuous expression if part of a copular construction, a light verb construction or a fixed expression. In addition, also named events such as “First World War” can have a multi-token extent. In all the other cases, the noun itself should be annotated alone, without including determiners or adjectives.

(30) *both in peace and war*= 2 annotations

²A list of English copulae is available on Wikipedia: https://en.wikipedia.org/wiki/List_of_English_copulae.

5.1. EVENT ANNOTATION GUIDELINES

5.1.3 Semantic Classes

Each annotated event mention should be classified by assigning a value to the **CLASS** attribute. The classification we have designed is based on semantic criteria and developed by re-elaborating the semantic categories of the Historical Thesaurus of the Oxford English Dictionary (HTOED) [Kay et al., 2009a] . The HTOED has been defined over several decades with the aim of conceptualising and classifying the meaning of the English language: it consists of a hierarchical structure with a primary tripartite division (*External World*, *Mental World*, and *Social World*), 37 categories and 377 sub-categories³.

In HTOED a distinction is made between categories connected with a physical existence and those having a social dimension: due to this subtle difference an event of movement can belong to the **TRAVEL AND TRAVELING**, the **SPACE** or the **MOVEMENT** category. In other words, discerning between physical and social dimensions is ambiguous. Therefore, starting from the original complex and extremely fine-grained classification, we worked to find an appropriate level of granularity by merging categories with a common conceptual core. This choice led us to create a unique class for events related to the concept of space (**SPACE-MOVEMENT**) and for those involving forces beyond scientific understanding or the laws of nature (**RELIGION-SUPERNATURAL**). In addition, we collapsed into the same class events in the area of production and trade of services and goods (**ECONOMY**), those in the public domain (**LAW-AUTHORITY**), and those involving all the types of living things and their health conditions (**LIFE-HEALTH**). Events connected to the faculties of the mind characterized by reasoning or knowledge are brought together in the **MENTAL-ABSTRACT** class, while instinctive or intuitive mental activities accompanied by a certain degree

³<http://historicalthesaurus.arts.gla.ac.uk/>

5.1. EVENT ANNOTATION GUIDELINES

of pleasure or displeasure are joined in the **EMOTIONS-EVALUATIONS** class. Figure 5.1 provides a graphical representation of how our classes were defined starting from the HTOED categories.

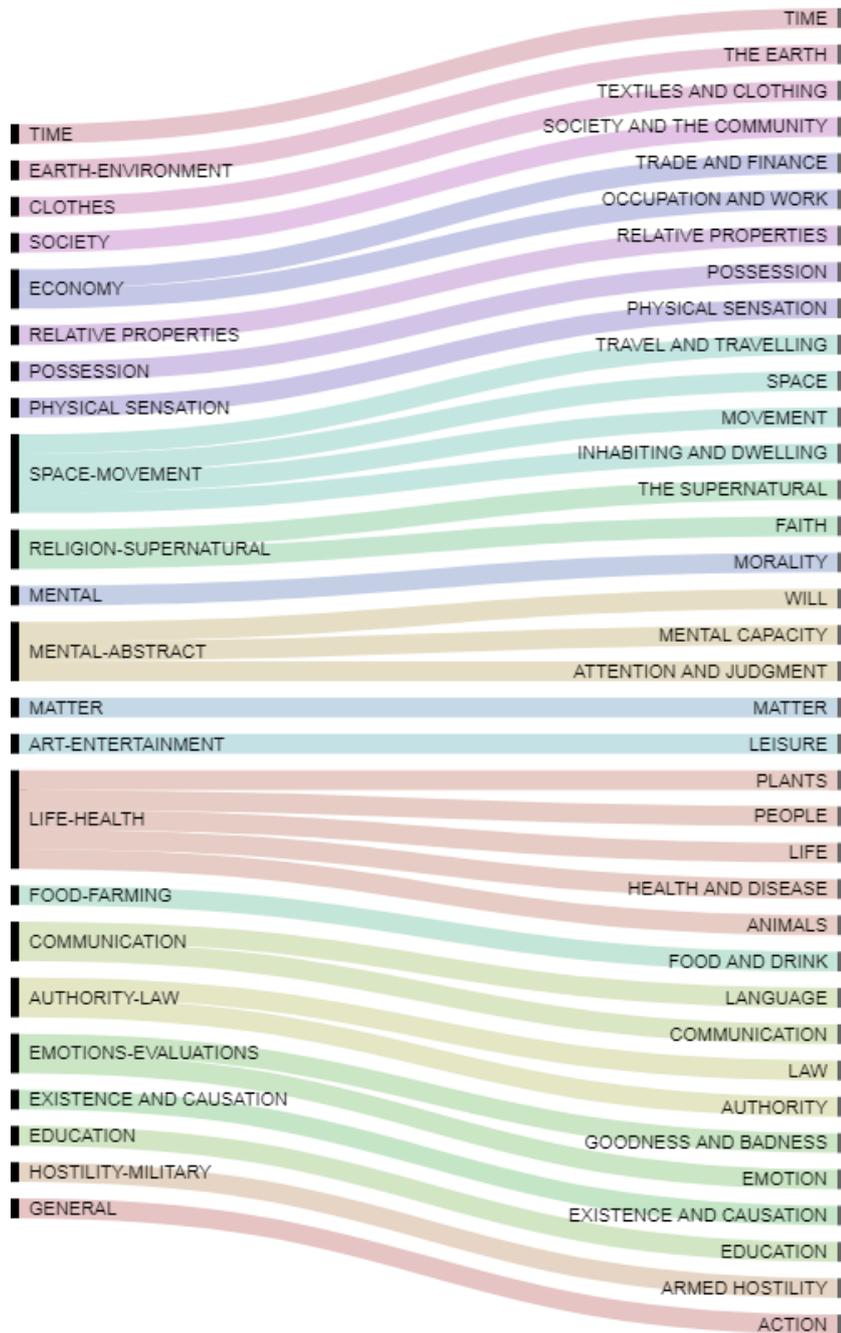


Figure 5.1: Mapping of our HISTO classes (left) to the second-level HTOED categories (right). Image created with RAWGraphs [Mauri et al., 2017].

5.1. EVENT ANNOTATION GUIDELINES

Description of Semantic Classes

Classes are described below together with a set of examples. Event extension is highlighted in bold.

1. EARTH-ENVIRONMENT, eventualities related to geography (31, 32), climate/weather conditions (33), environmental issues (34).

(31) *the streets **are like** caverns*

(32) *the settings **are** Spanish*

(33) *It has been **raining** for days*

(34) ***deforestation** has denuded the mountain-side*

2. LIFE-HEALTH, eventualities related to living things, i.e. humans (35), animals and plants (36), including life (37), death (38), physical conditions, diseases and medical treatments (39).

(35) *he **was a** Caprian paesant*

(36) *oranges do not **grow up***

(37) *which **lives upon** public charity*

(38) *he will not **kill** the woman*

(39) *in charges of contagious **diseases***

3. FOOD-FARMING, eventualities pertaining to food (40), food preparation and consumption (41), drink (42), agriculture (43) and hunting (44).

(40) *they **are only** nuts*

(41) *let us **breakfast** together*

(42) *I luxuriously **sip** my coffee*

(43) *an elderly man was **plowing** with a pair of oxen*

5.1. EVENT ANNOTATION GUIDELINES

(44) *this Diana **is not a huntress***

4. CLOTHES, eventualities associated to textiles (45), clothes (46, 47) and other personal belongings (48).

(45) *they are renowned for their skill in **weaving***

(46) *he **took off** his hat*

(47) *he **wore** only a shirt*

(48) *it **is a rather heavy portmanteau***

5. MATTER, eventualities connected to substances and materials, their properties, constitution and conditions (49). This class includes terms relating to liquids (50), solids, gases, electricity (51), light (52), colours (52), shapes.

(49) *burdens that **seem too heavy***

(50) *the miracle of **liquefaction***

(51) *the power **transmitted** need not be necessarily destructive*

(52) *their full black hair **shines** like satin*

(53) *his faces **being rather red***

6. EXISTENCE-CAUSATION, eventualities relating to the concepts of being as in existential clauses [McNally, 1998] (54), occurring (55), existing and causation (56) and their lack. It includes both creation (57) and destruction (58), damage (59), break and demolition.

(54) *in this court **are** a number of handsome sarcophagi*

(55) *these **occurrences** are fanning a spirit of revenge*

(56) *the cases **caused** me a genuine thrill*

(57) *a cloud is **occasioned** by the column of steam*

5.1. EVENT ANNOTATION GUIDELINES

(58) *sudden **destruction** of the buried city*

(59) *the regular ambulance was **wrecked** last night*

7. SPACE–MOVEMENT, brings together all the eventualities pertaining to space (extensions (60), directions, presence (61)), movements (motions of entire bodies (62) and part of bodies, objects (63) but also changes of place, transfers, impacts), lack or end of movement (various stages of inactivity, such as stillness (64), stops (65), waiting), and travel (referring to ways of travelling on land (66), by water (67), by air).

(60) *mitre of gold is **covered** with precious gold*

(61) *is there any significance in the **presence** of the Mayor?*

(62) *he had **walked through** the shaded park*

(63) *the waters of two fountains **mingle** and **flow** together*

(64) *the Temple of Minerva **standing** beside twelfth-century buildings*

(65) *the little inn at which we are **stopping***

(66) *after our exciting **drive***

(67) *we **sailed** from New York six weeks ago*

8. TIME, eventualities associated to frequency (68), duration (69), change (70), age (71) and the spending of time (72). Aspectual terms (73) are included in this class because they denote distinct parts of the internal temporal structure of eventualities [Damova and Bergler, 2000].

(68) *this **is the first time** the cup will leave France*

(69) *the raid **lasted** for about half an hour*

(70) *abrupt **changes** of temperature*

(71) *I'm **only 20 years old***

5.1. EVENT ANNOTATION GUIDELINES

(72) *she **spent** some weeks at Sydney*

(73) *for the **beginning** of our drive*

9. **GENERAL**, general eventualities denoting not specific operations upon something like doing, using (74), trying (75), helping (76), finding but also events and states relating to safety/danger (77), difficult/easiness, success/failure.

(74) *the very best tobacco **used** in the cigar factories*

(75) *a man who is **trying** to free himself*

(76) *with the **help** of a band*

(77) *Doctor Antonio hesitated about **imperilling** her neck*

10. **RELATIVE PROPERTIES**, eventualities pertaining to measurements (78, 79), numbers (80) (except those relative to the temporal dimension that have to be annotated with the **TIME** class) and quantities (81).

(78) *this island of Luzon **is so large***

(79) *he **counted** the guns which were fired*

(80) *because of a **reduction** in their wages*

(81) *these Paris delegates **are thirty-five***

11. **RELIGION-SUPERNATURAL**, eventualities relating to religions (government of organized religions (82), sects, religious ceremonies (83) and worship in general) and the supernatural (84) (occult, paranormal, supernatural manifestations, deities).

(82) *Mr. Zeglen **is a young priest***

(83) *the **high mass** is celebrated*

(84) *the departed **haunt** the silent town*

5.1. EVENT ANNOTATION GUIDELINES

12. **MENTAL-ABSTRACT**, includes all mental actions and processes (85, 86) (reasoning, thinking, believing, knowing, understanding, remembering), attention and judgment (87), expressions of will (88) (necessities, inclinations, intentions, decisions and motivations). The lack of mental capacities, attention and free will is included as well in this class (89).

(85) *Victor Emanuel **seems** to have **thought** that...*

(86) *Sir George **had knowledge** of the traditions*

(87) *he could **take care** of me and himself*

(88) *having **decided** to meet Zelfhine and Angela*

(89) *people **were fairly ignorant***

13. **EMOTIONS-EVALUATIONS**, emotional actions, states and processes or eventualities expressing the lack of emotions (90, 91) (excitement/calmness, pleasure/suffering, compassion/indifference, courage/fear, love/hate). This class includes also other aspects of subjectivity, i.e. evaluations about goodness or badness, inferiority or importance (92).

(90) *the days brought me **enjoyment** and **delight***

(91) *the witnesses **were amazed** at the man's **calmness***

(92) *the second election **was not less important** that the first*

14. **POSSESSION**, includes eventualities associated to concepts such as having, not having, losing, taking, giving, allocating, acquiring, receiving, sharing (93, 94, 95, 96) and the opposition between wealthy and poverty (97).

(93) ***having** like it four colossal bronze lions at the base*

(94) *my mother is **keeping** it*

5.1. EVENT ANNOTATION GUIDELINES

(95) *they would not **take** a large sum of money for the experience*

(96) *its **withdrawal** or **supply** will bring about the same results*

(97) *all his neighbors would testify to his **poverty***

15. **COMMUNICATION**, linguistic actions, states and processes that is eventualities connected to both the intellectual activity of speaking a language, naming things, producing speech acts (98, 99, 100) and the social activity of expressing, transmitting and receiving information in different ways through the media (101, 102) (e.g., writing, printing and publishing books, reading books and journals, using telecommunication technologies or correspondence).

(98) *a lovely little square **called** Acquaverde*

(99) ***crying out**: “ecco, ecco, signora!”*

(100) *she **refused** to be blindfolded*

(101) *any one who can **write** letters as interesting as yours*

(102) *your mother will remember **reading** this story to me*

16. **SOCIAL**, eventualities involving the society in general or a specific community. The class includes social actions, states and processes such as the participation, or lack of participation (103), in meetings (104) and relationships of different types: intimate, between family members (105), within groups (106) and associations.

(103) *spend a summer month in much-advertised **seclusion***

(104) *the **meeting** was addressed by anarchists*

(105) *after her **marriage** with Lord Cleverton*

(106) *she will be calling me soon to **join** her*

5.1. EVENT ANNOTATION GUIDELINES

17. **HOSTILITY-MILITARY**, eventualities related to different aspect of military life (107, 108) (operations, service, use of weapons) and acts of hostility. War and peace, attack and defense are included (109, 110).

(107) *they shall not **conscript***

(108) *Tully gave this story of the **shooting***

(109) *America stands supremely for **peace***

(110) *this chapel has escaped the vicissitudes of **revolutions** and **wars***

18. **AUTHORITY-LAW**, eventualities associated to political and governmental activities (111) and in general to the exercise (112) or lack of authority (power, rule) but also to criminal activities (113) and to the legal system (114, 115) (legislation, legal power, punishments). Among criminal activities, offences against the person such as murder, manslaughter and wounding are to be annotated as **LIFE-HEALTH** while offences against the property like theft and robbery fall in the **POSSESSION** class.

(111) *favorite candidate for the next municipal **election***

(112) *Commander Clifford **commanding** the Pampanga*

(113) *they had willfully **disobeyed** the law and were **locked up***

(114) *during the **trial** here in Buffalo*

(115) *the troops were already drawn up for the **execution***

19. **EDUCATION**, eventualities pertaining to teaching, learning but also to the administration of educational institutions (116, 117).

(116) *he was **graduated** from Princeton University in 1906*

(117) *they are not learning anything*

5.1. EVENT ANNOTATION GUIDELINES

20. **ECONOMY**, eventualities connected to money (118) (change of money, payments and taxation), commerce (119) (business affairs, trading operations, buying and selling), work and employment (120) (occupations but also lack of work).

(118) *she **sold** her pearls to **raise money** to feed the poor*

(119) *conditions will quickly settle and **trade** revive*

(120) *Larcher **is a locksmith***

21. **ENTERTAINMENT-ART** eventualities related to entertainment (121) (night-life, hobbies), arts (122, 123) (performing art, music, visual arts), sports (124) and games in general (125).

(121) *to expiate his **gambling** debts*

(122) *not content with this they **gave a dance** that same evening*

(123) *this **is a curious statue***

(124) *I **wrestled** quite a bit*

(125) *the number of **games** to be **played** here will be at least three*

22. **PHYSICAL SENSATION**, eventualities related to the perception by senses (126, 127) (touch, taste, smell, sight, hearing) but also the sleeping/waking (128) and the cleanness/dirtiness (129). The use of cigarette or drugs is included in this class as well (143).

(126) *the wind is scarce **felt**, though you may **hear** it **sighing***

(127) *an unexpected **glimpse** into the valley*

(128) *cancel all speaking engagements and **take a complete rest***

(129) *the Neapolitan city **is even dirtier***

(130) *students drank, talked and **smoked***

5.1. EVENT ANNOTATION GUIDELINES

How to Assign the Class Value

In light verb constructions and constructions with support or predicational copulae, the main meaning resides not in the head verb but in the noun or copula complement: for this reason, the class attribution is based on the meaning of the noun or adjective attached to the verb.

(131) *if we **make the trip** in an automobile* = SPACE-MOVEMENT

(132) *war-horses **were monsters*** = RELIGION-SUPERNATURAL

(133) *they **were capable*** = MENTAL-ABSTRACT

(134) *the raisers **were less numerous*** = RELATIVE PROPERTIES

As for phrasal verbs and fixed expressions, the meaning of the whole linguistic unit, and not of the single parts, should be taken into consideration to assign the correct class.

(135) *how three women **get on** together* = SOCIAL

(136) ***get back** to Chiaia by five o'clock* = SPACE-MOVEMENT

Some of the previous examples show that the same token (like “were” and “get”) can be annotated with different classes according to the context, i.e. the semantics of the other parts that make up the whole event mention. In addition, other classes could be attributed to that token when it appears outside a light verb or copular construction. The verb “to be”, for example, is annotated as belonging to the EXISTENCE-CAUSATION class when in existential clauses, often in combination with the word “there” used as pronoun, while the verb “to get” can be assigned to the POSSESSION class.

(137) *there **is** a handsome modern statue* = EXISTENCE-CAUSATION

(138) *I haven't **got** it now* = POSSESSION

5.2. DATASET CONSTRUCTION

The verb “to fall” serves as another example of the importance of the context for class attribution. It is usually annotated as belonging to the SPACE-MOVEMENT class, both as a single-token verb and as part of a phrasal verb, but when referring to the weather, the correct class is the ENVIRONMENT one.

(139) *falling down with a rattling noise* = SPACE-MOVEMENT

(140) *two bombs fell in the Thames* = SPACE-MOVEMENT

(141) *the rain fell in torrents* = EARTH-ENVIRONMENT

Similarly, “to break” usually refers to the separation of objects into pieces and is annotated with the EXISTENCE-CAUSATION class; however, when referring to an injury involving the fracture of a body part, the class is LIFE-HEALTH

(142) *the concussion of this bomb broke glass* = EXISTENCE-CAUSATION

(143) *suffered a broken leg* = LIFE-HEALTH

5.2 Dataset Construction

We applied the guidelines described in Section 5.1 to a newly created collection of historical texts named *Histo Corpus*. The following subsections describe the corpus and its annotation process with details on inter-annotator agreement.

5.2.1 Corpus Description

The *Histo Corpus* (henceforth, HC) consists of historical texts of two different genres, namely travel narratives and news, published between the second half of the XIX Century and the beginning of the XX Century.

5.2. DATASET CONSTRUCTION

News have been taken from the newspaper portal of Wikisource⁴, the Wikimedia Foundation website containing a digital library of source text transcriptions free of copyright. We selected news covering various topics, such as murders, conflicts, sports, movie reviews, obituaries, scientific discoveries and gossip on celebrities. The choice of news as a genre to be included in the corpus is in line with past and current trends of annotated corpora as shown in our analysis reported in Chapter 2. However, the historical nature of the texts and the diversity of topics covered by the news makes them particularly interesting for annotation.

On the other hand, travel narratives are not much explored in computational linguistics: exceptions are the ANC (American National Corpus) and GUM (Georgetown University Multilayer) corpora [Ide and Macleod, 2001, Zeldes, 2017] that, however, contain only contemporary texts, and the collection of historical German travel guides developed within the *traveldigital* project⁵ [Czeitschner and Krautgartner, 2017]. Nevertheless, to the best of our knowledge, no corpus of travel narratives with event annotation has been released before⁶. We choose this particular genre because travel writings are powerful sources of information for many research areas, such as art history, ethnography, geography and cultural history [Burke, 1997]. Thus the automatic extraction of events from them can be useful for researchers in several domains of the Humanities.

Travel narratives included in *HT* have been extracted from a larger collection of texts we have created with the aim of fostering research on travel writings with digital and computational methods [Sprugnoli et al., 2017c, Sprugnoli, 2018]. More specifically, this collection consists of 57 books, for a total of 3,630,781 tokens: all the books are available in a cleaned text

⁴<https://en.wikisource.org/>

⁵<https://traveldigital.acdh.oeaw.ac.at/>

⁶Travel guides in the *traveldigital* project are annotated following a domain-specific thesaurus that includes a very limited type of events, that is tourist activities such as excursions and carriage rides.

5.2. DATASET CONSTRUCTION

format and thirty of them are also distributed in TEI-XML on a dedicated website⁷. These books, both travel narratives (reports, diaries, collections of letters) and travel guides, were taken from Project Gutenberg⁸, are about Italy, were written by Anglo-American authors, and were published between the country’s unification (in 1861) and the beginning of the 1930’s. We choose this period because in the second half of the 19th Century, the tradition of the Grand Tour⁹ declined and leisure-oriented travel emerged. This radical transformation was enabled by technological, economic and sociological factors, such as the development of steam-powered ships and of the railway network, the growth of the Anglo-American economy and a greater emancipation of women that led to having more female travelers [Schriber, 1995]. Moreover, after Italian unification, new routes to Southern Italy and the islands were opened, so that travelers’ attention was no longer limited to the classic destinations in North and Central Italy, such as Venice, Florence and Rome [Ouditt and Polezzi, 2012]. Texts annotated in HT come from this large resource: all the documents included in the corpus are travel narratives, in particular reports and letters, and are about different Italian locations such as Naples, Genoa, Assisi and Viareggio.

Table 5.1 shows details on the number of documents and tokens in *HT* together with their period of publication. Even if HT is not as large as other corpora annotated with temporal information, at the moment of writing it is the largest available corpus annotated with events in the historical domain (see Table 2.4 for comparisons with the *ModeS Timebank* and the *De Gasperi Corpus*).

⁷<https://sites.google.com/view/travelwritingsonitaly>

⁸<https://www.gutenberg.org/>

⁹The Grand Tour was the traditional educational journey undertaken by upper class people, especially young men, that became customary among British nobles starting from the 17th century [Buzard, 2002]. The typical itinerary of the Grand Tour included several Italian cities appreciated for their heritage of ancient Roman monuments and Renaissance culture [Sweet, 2012].

5.2. DATASET CONSTRUCTION

	DOCS	TOKENS	PERIOD OF PUBLICATION
Travel Narratives	25	28,259	1865-1921
News	47	27,821	1883-1926
TOTAL	72	56,080	

Table 5.1: Statistics on the *Histo Corpus*

5.2.2 Corpus Annotation

The *Histo Corpus* has been annotated following the guidelines described in Section 5.1 and using the web-based CAT annotation tool [Bartalesi Lenzi et al., 2012]¹⁰. This subsection contains description and results of the inter-annotator agreement performed to check the soundness of the guidelines and the feasibility of the proposed tasks. Then we give details on the annotated data with an analysis of the main differences between events annotated in the two genres forming the *Histo Corpus*.

Inter-Annotator Agreement

We measured the inter-annotator agreement (IAA) [Artstein and Poesio, 2008] on a subset of the *Histo Corpus* balanced between the two genres in terms of token number: one travel narrative and four news about different topics (national and foreign policy, sport, scientific discoveries) were selected for a total of 1,200 tokens. Two annotators performed the work independently using the guidelines reported in Section 5.1: one was the author of the thesis, the other was non involved in the development of the guidelines. Both annotators were not English native speakers but were expert in linguistic annotations.

Results of the IAA are reported below with different metrics. The Dice coefficient [Dice, 1945] is given for the identification of event mentions distinguishing between the agreement calculated on extensions perfectly

¹⁰CAT output format will be described in Section 6.1

5.2. DATASET CONSTRUCTION

detected by both the annotators and the one measured on the number of annotated tokens shared by both annotators, thus considering also a partial match. In other words, with the Dice Coefficient we measure the agreement in determining whether each token is or is not part of an event mention. In addition, we provide the Cohen’s kappa [Cohen, 1960] so to also measure the pairwise agreement taking into consideration agreement that would be obtained by chance. For event classification, we calculated both the overall accuracy and the Cohen’s kappa on mentions detected by both annotators:

- EVENT MENTION DETECTION:
 - Dice Coefficient macro-average at tag level (perfect match): 0.85
 - Dice Coefficient macro-average at token level: 0.87
 - Cohen’s kappa: 0.85

- EVENT CLASSIFICATION:
 - Accuracy: 0.74
 - Cohen’s kappa: 0.71

In computational linguistics, a kappa score of 0.80 is considered a threshold to exceed for having data with good reliability [Landis and Koch, 1977, Carletta, 1996]. Our results on event mention detection are particularly good given the presence of multi-token and discontinuous mentions: the agreement on perfect match is only slightly lower than the one at token level (0.85 *vs* 0.87) meaning that mentions can be detected in a consistent way. Disagreements were due to differences in the inclusion of prepositions in the event extent (“twister over”) and to the non-identification of copular constructions (“the average speed was 44 miles per hour”). Another problematic case is given by polysemous event nominals like “story” in the

5.2. DATASET CONSTRUCTION

following sentence, which may denote both an event and an information object: “a witness of the truth of the story”.

As for event classification, results are lower but still satisfactory given the complexity of the task with 22 different options. Seven out of 22 classes achieved a perfect agreement: **COMMUNICATION**, **EDUCATION**, **FOOD-FARMING**, **LIFE-HEALTH**, **PHYSICAL SENSATIONS**, **SOCIAL** and **ECONOMY**. Disagreement in the other classes was registered, for example, for cases of figurative uses of verbs (e.g. “the white cap of Vesuvius worn generally like the caps of the Neapolitans”). Moreover, annotators tended to overuse the class **GENERAL** as a backup category in case of uncertainties.

By comparing the IAA on the *Histo Corpus* with the agreement reported for other schemes dealing with event annotation, it is worth noticing higher results both for the extent and the class of event mentions. In TimeBank 1.2, the agreement is of 0.81 on partial match, 0.71 on perfect match and 0.67 on class assignment¹¹. For the data used in the Event Nugget task in 2015, the agreement on event detection does not reach 0.80 and is below 0.70 on event classification [Song et al., 2016].

Annotated Data

Table 5.2 reports the number of annotated events in HT per class and text genre. News and travel narratives show, for almost all the event classes, a statistically significant difference (at $p < 0.05$ and calculated with the z test¹²) in their distribution.

The high occurrence of events belonging to the **SPACE-MOVEMENT** class in both genres is due to the broad definition of the class that covers the three main concepts of motion [Sablayrolles, 1995], i.e. locations, positions

¹¹Data reported in the TimeBank 1.2 documentation: <http://www.timeml.org/timebank/documentation-1.2.html>

¹²The z test is a parametric statistical test used to verify if the mean value of a distribution differs significantly from a certain reference value [Sprinthall, 2003].

5.2. DATASET CONSTRUCTION

and postures, and both factive and fictive motions. Examples of a change of location (144), position (145) and posture (146) are given below. These are cases of factive motions, while examples 147 and 148 contain events of fictive motions, that is “linguistic instances that depict motion with no physical occurrence” [Talmy, 1996]:

(144) *Marcel Renault **arrived** first*

(145) *the pigs used to **run about** in the principal streets of Naples*

(146) *a man **lay** in one of the entrances to the Union Station*

(147) *lemon trees **covered** with ripening fruit*

(148) *a deep ravine **surrounded** by mountains*

Also the range of COMMUNICATION events is wide and particularly relevant in the news that typically report the testimonies of observers and witness of what is recounted (see examples 149 and 150):

(149) *he **told** Inspector Fairey*

(150) *he **admitted** that the council may have made mistakes*

The predominance of LIFE-HEALTH, HOSTILITY-MILITARY and AUTHORITY-LAW is characteristic of news only: these classes cover events expressing, among others, murders and injuries (151), local riots (152), international war offences (153), public administration and judicial process (154, 158) therefore they are particularly frequent in news about crimes, conflicts and politics.

(151) *fifteen persons were **killed** and seventy **injured***

(152) *he just **hauled off** and **hit** me*

(153) *the **siege** of Mafeking*

5.2. DATASET CONSTRUCTION

CLASS	NEWS	TRAVEL	TOTAL
SPACE-MOVEMENT*	791	963	1,754
COMMUNICATION*	571	377	948
GENERAL*	516	315	831
MENTAL-ABSTRACT	420	419	839
EMOTIONS-EVALUATIONS*	239	450	689
EXISTENCE-CAUSATION*	360	296	656
PHYSICAL SENSATIONS*	200	324	524
LIFE-HEALTH*	215	144	359
POSSESSION	173	166	339
HOSTILITY-MILITARY*	260	25	285
TIME	119	120	239
AUTHORITY-LAW*	205	9	214
ENTERTAINMENT-ART*	103	68	171
ECONOMY*	115	46	161
RELATIVE PROPERTIES	67	67	134
SOCIAL*	96	32	128
MATTER*	37	86	123
ENVIRONMENT*	23	71	94
FOOD-FARMING*	13	56	69
CLOTHES	37	21	58
RELIGION-SUPERNATURAL*	2	27	29
EDUCATION	16	7	23
TOTAL*	4,578	4,089	8,667

Table 5.2: Annotated events per class and text genre together with the total amount of annotations. The asterisk indicates whether the class has a statistically significant difference in the distribution over the two genres.

(154) *Czolgosz was **sentenced** to die*

(155) *the **appointment** of Captain George Sitwell*

On the contrary, the PHYSICAL SENSATIONS class is strongly represented in travel narratives, in which the writer reports her/his experiences with local people and local environments:

5.2. DATASET CONSTRUCTION

(156) *we **see** no pretty young Genoese women*

(157) *one **perceives** Nature has rejoiced in her work there*

(158) *what joy to **listen** to analphabets for a change*

As for the extent, 897 event mentions in the news and 860 in travel narratives are annotated with a multi-token span: these numbers correspond to the 19.6% and the 21.4% of the total number of events in the two genres respectively. This difference is not statistically significant at $p < 5$. The majority of multi-token events are copular constructions with the verb “to be” (45.5%) but other verbs used as copulae are present as well, for example “to become” and “to fell”. These constructions are mainly annotated with the class **EMOTIONS-EVALUATIONS** while the second most common class for multi-token event mentions is **SPACE-MOVEMENT**. This class covers many phrasal verbs such as “go out”, “go away”.

The last row of Table 5.2 shows that also the difference in the total number of annotated events in news and travel narratives is statistically significant with the former having a higher occurrence of event mentions (4,578 *versus* 4,089). We have investigated this phenomenon by connecting the analysis of annotated events with the notion of Content Types we have developed during the PhD [Sprugnoli et al., 2017a].

5.2.3 Interlinking Events and Content Types

Content Types (henceforth CTs) are text passages with specific semantic and functional roles: they contribute to the overall message or purpose of a text and make explicit the functional role of a discourse segment with respect to its content, i.e. meaning.

Over the years, different typologies have been proposed to classify whole texts [Werlich, 1976, Biber, 1989, Chatman, 1990, Adam, 1985, Longacre,

5.2. DATASET CONSTRUCTION

2013] or text passages. Several annotation schemes, often based on genre-specific taxonomies, have been also developed. This is the case, for example, of the detection of the main components of scientific discourse in scholarly publications [Teufel et al., 2009, Liakata et al., 2012, De Waard and Maat, 2012, Burns et al., 2016]. Also the annotation of content zones, i.e., functional constituents of texts, is genre-dependent thus different schemes have been developed to address the phenomenon in movie reviews, legal documents and news [Bieler et al., 2007, Stede and Kuhn, 2009, Baiamonte et al., 2016]. In our work, we have instead identified seven classes of CTs, five of which are inspired by Werlich’s typology, while two (**OTHER** and **NONE**) were introduced in our scheme to account for undefined or unclear cases. We have chosen these seven classes because they provide a good level of generalization for characterizing documents with different structures (e.g. news articles *versus* scientific article), and can be applied across different domains and genres.

Classes are identified at clause level because different portions of the same sentence can be characterized by different CTs. A brief description of each class is given below together with an example (the symbol “//” marks clause boundary) :

- **NARRATIVE**: clauses containing events and states that can be anchored to a hypothetical timeline even if not reported in a perfect sequential order, due for example to flashbacks; e.g., *Bombs were dropped at one coast town, // three women being slightly injured.*
- **EVALUATIVE**: clauses with explicit evaluation markers; e.g., *The offer which Telerate’s two independent directors have rejected as inadequate.*
- **DESCRIPTIVE**: clauses presenting tangible and intangible characteristics of entities, such as objects, persons or locations, thus creating a mental picture of these entities in the readers mind; e.g., *The road*

5.2. DATASET CONSTRUCTION

winds above, beneath, and beside rugged cliffs of great height.

- **EXPOSITORY**: clauses expressing generalizations with respect to a class.; e.g., *All Italians are dandies.*
- **INSTRUCTIVE**: clauses expressing procedural information; e.g., *At last you cross that big road // and strike the limestone rock.*
- **OTHER**: clauses containing text in foreign languages, phatic expressions, references to the reader; e.g., *Madame est servie.*
- **NONE**: clauses that cannot be labeled with any of the previous classes, such as the date at the beginning of a news e.g., *May 19, 1917*

Text contained in the *Histo Corpus* have been annotated with CTs to gain insight into the function and the semantics of their content. Table 5.3 reports quantitative results of this annotation, that is the number of annotated CTs per genre and the results of the inter-annotator agreement calculated over a subset of the corpus. Also in this case, the two genres show a difference in CTs distribution that is statistically significant (at $p < 0.05$ and calculated with the z test).

News, as expected, are characterized by a large amount of **NARRATIVE** CTs that cover the 74.7% of all the CTs annotated in that section of the corpus. **NARRATIVE** CTs are about events taking place, their purpose is to tell a story and this explains the greater presence of events in news with respect to travel narratives. On the contrary, **EVALUATIVE** and **DESCRIPTIVE** CTs are peculiar of travel narratives. The former contain opinions and personal feelings travellers express in their writings about places and people met, thus they are strictly related to the **EMOTIONS-EVALUATIONS** events, whose number is significant in travel narratives. For instance, example (159) is made by a **NARRATIVE** CT, containing a **COMMUNICATION** event

5.3. CHAPTER SUMMARY

CONTENT TYPE	NEWS		TRAVEL	
	#	k	#	k
NARRATIVE*	1,993	0.89	1,741	0.88
EVALUATIVE*	324	0.94	627	0.90
DESCRIPTIVE*	169	0.76	476	0.86
EXPOSITORY*	10	0.81	86	0.93
INSTRUCTIVE*	25	-	7	0.65
OTHER*	30	-	194	0.92
NONE*	116	1	19	1
TOTAL	2,667		3,150	

Table 5.3: Number (#) of annotated CT in the *Histo Corpus* and Cohen’s kappa (k), calculated between two annotators, for each type of CT. An asterisk marks that in all the cases there is a statistically significant differences in the distribution of CTs over the two genres.

(“tell”), followed by an **EVALUATIVE** CT containing an event of the class **EMOTIONS-EVALUATIONS** (“is delightful”):

(159) *Mrs. Coxe tells us // is delightful*

As for **DESCRIPTIVE** CTs, they convey word pictures of what travellers saw: their high occurrence in travel narratives is connected with the high occurrence of events in the classes **MEASURE** (160), **MATTER** (161) and **ENVIRONMENT** (162) that provide an analysis of the visited places as experienced during the journey:

(160) *the church of St. Severo is full of fine modern statues*

(161) *the delicate mauve pink of her towers glowed with a rosy hue*

(162) *the country is so hilly*

5.3 Chapter Summary

In this Chapter we present new annotation guidelines designed to meet historians’ requirements about event mention detection and classification

5.3. CHAPTER SUMMARY

as emerged from the online questionnaire described in Chapter 4. Events are defined following Bach’s notion of eventuality, thus taking into consideration all types of actions, processes and states. Event detection is based on different parts of speech and syntactic criteria so to include in the annotation several constructions and also multi-token events both in continuous and discontinuous textual sequences. As for the classification, we adopted a semantic approach choosing a comprehensive set of 22 classes.

Another contribution of this Chapter is given by the *Histo Corpus*, a new resource consisting of historical texts manually annotated with event mentions following our guidelines. This corpus contains texts of two genres, namely news and travel narratives: the latter constitute an under-investigated genre in NLP and, more specifically, is a novel type of text in the domain of temporal information extraction.

The inter-annotator agreement calculated on the dual annotation of a subset of the *Histo Corpus* proved the soundness of the guidelines both in terms of event detection and event classification. For the former we registered a kappa of 0.85 and for the second a kappa of 0.71: both results are in line or even higher than the ones reported for other event annotation schemes, considering the same tasks.

The outcome of the manual annotation has been discussed according to various aspects: we calculated whether the difference in event distribution was statistically significant over the two text genres, we analysed the predominance of different event classes in news and travel narratives and we connected event annotation to the notion of Content Types.

In the following Chapter, we report on experiments conducted using the *Histo Corpus* to train and test systems for automatic event detection and classification using our newly introduced annotation scheme. More specifically, we developed two classifiers, one for only mention detection and one for also event classification, adopting the Conditional random fields (CRFs)

5.3. CHAPTER SUMMARY

modeling method and by applying a neural architecture (biLSTM).

5.3. CHAPTER SUMMARY

Chapter 6

Events in Historical Texts: Automatic Annotation

After having defined annotation guidelines and manually tagged a corpus accordingly, as described in the previous Chapter, here we report on experiments on the automatic detection and classification of event mentions.

Experiments were carried out using the annotated *Histo Corpus* divided in a training, a test and a dev set (Section 6.1). Then, we followed two different approaches. On one hand, we implemented two CRF classifiers detailed in Section 6.2: one is aimed at identifying the correct span of event mentions and the other at assigning the correct class to each event mention starting from raw text. This last task implies the identification of mentions: in other words, no golden event mentions are given in input to the system. For the CRF classifiers we provide an analysis of features and of the impact of different context windows on the precision, recall and F1-score. On the other hand, we used a BiLSTM implementation for sequence tagging: also in this case both tasks, event detection and event classification, were taken into account. This implementation [Reimers and Gurevych, 2017a] does not require any feature engineering: it is based on a neural architecture and on the use of dense vectors representing words. In Section 6.3 we describe the general architecture of the system and the

6.1. DATA PREPARATION

results obtained by evaluating different hyperparameters' options and pre-trained word embeddings.

<pre><Document doc_name="file.txt"> <token t_id="1" sentence="0" number="0">we</token> <token t_id="2" sentence="0" number="1">set</token> <token t_id="3" sentence="0" number="2">forth</token> <token t_id="4" sentence="0" number="3">at</token> <token t_id="5" sentence="0" number="4">eight</token> <token t_id="6" sentence="0" number="5">o'clock</token> <token t_id="7" sentence="0" number="6">.</token> <Markables> <EVENT_MENTION m_id="1" comment="" class="SPACE-MOUMENT"> <token_anchor t_id="2"/> <token_anchor t_id="3"/> </EVENT_MENTION> </Markables> <Relations> </Relations> </Document></pre>	<pre>MENTION DETECTION ONLY TASK we 0 set B-EVENT_MENTION forth I-EVENT_MENTION at 0 eight 0 o'clock 0 . DETECTION+CLASSIFICATION TASK we 0 set B-SPACE_MOVEMENT forth I-SPACE_MOVEMENT at 0 eight 0 o'clock 0 .</pre>
---	---

Figure 6.1: Example of a file in the CAT XML format (left) and in the corresponding converted BIO/IOB2 notation (right) for the two tasks.

6.1 Data Preparation

As a first step, we automatically converted annotated files from the CAT format to the BIO/IOB2 notation. The former is the stand-off XML format of the CAT annotation tool: in it, different annotation layers are contained in separate document sections and related to each other and to the source text through pointers. The latter is a tagging scheme in which a “B-” tag marks the first token of an annotated segment (in our case a segment is an event mention), “I-” is used for all the other tokens within the span of the same segment and “O-” marks tokens that do not belong to the segment [Sang and Veenstra, 1999]. We chose the BIO/IOB2 notation because for the biLSTM (bidirectional Long-Short Term Memory) architecture it has proved to perform better than other notations such as IOB1 [Reimers and Gurevych, 2017a] in which the “B-” tag marks the beginning of an annotated segment only when it immediately follows another annotated

6.2. CRF CLASSIFIERS

segment.

Figure 6.1 shows an example of CAT and BIO/IOB2 formats. For the mention detection task, the “B-EVENT_MENTION” and “I-EVENT_MENTION” tags are used to indicate the span of each event while for the classification task, tags are used to specify the event class and, implicitly its extension.

After the conversion, we divided the *Histo Corpus* in a training and a test set for the CRF models and in a training, test and development set for the neural architecture. Given that we wanted to compare the performances of the two approaches, we used the same training (80% of the whole corpus) and test (10%) sets: the remaining 10% of the data was used as development set for the biLSTM system. The files were chosen randomly as for class value but we balanced the distribution in each section across the two genres.

6.2 CRF Classifiers

For the first set of experiments, we implemented two linear CRF classifiers using CRFSuite [Okazaki, 2007], a software for labeling sequential data: it contains different state-of-the-art training methods and an integrated evaluation functionality to compute Precision, Recall and F1-score on test data¹. In all the experiments, we used the default training algorithm of CRFSuite (L-BFGS, Limited-memory Broyden-Fletcher-Goldfarb-Shanno) with L1 regularization. In addition, we put a threshold to ignore features whose frequency of occurrence in the training data is below 2 and made CRFSuite generate both state and transition features.

As for features, we chose a simple set of three beyond the token itself: (i) lemma, (ii) PoS and (iii) text genre. The first two were extracted by

¹<http://www.chokkan.org/software/crfsuite/>

6.2. CRF CLASSIFIERS

processing the texts in *Histo Corpus* with Stanford CoreNLP [Manning et al., 2014] and the third marks the opposition between news and travel narratives at document level. To better evaluate the performance of our models, we developed a baseline system using only tokens as features.

6.2.1 Evaluation

In the following subsections we present the results of several evaluations carried out on the test set: in particular, we analyse the impact of the features and of the size of the context window on the performance of the classifiers.

Feature Analysis

To analyse how the different features influence the overall performance of the classifiers, we tested the models with all the features and then removed them one by one. The results of this evaluation for the task of event mention detection are reported in Table 6.1 while Table 6.2 shows the results for the classification of events, a task that implies the identification of mentions because no gold mentions are given in input. Tables provide information about the macro-average precision (P), recall (R) and F1 calculated considering a context window of $[+/-2]$.

<i>MENTION DETECTION ONLY</i>	P	R	F1
ALL FEATURES	84.93%	82.13%	83.44%
- without lemma	85.73%	81.89%	83.69%
- without PoS	81.32%	77.00%	78.88%
- without genre	84.72%	81.91%	83.22%
- with PoS and genre	85.93%	83.04%	84.38%
BASELINE (only tokens)	80.35%	74.64%	77.14%

Table 6.1: Performance, in terms of precision (P), recall (R) and F1, of the CRF model for event mention detection with different settings of features.

6.2. CRF CLASSIFIERS

<i>DETECTION+CLASSIFICATION</i>	P	R	F1
ALL FEATURES	30.39%	25.89%	26.03%
- without lemma	27.32%	21.59%	22.18%
- without PoS	32.24%	23.96%	25.62%
- without genre	31.96%	26.20%	26.00%
- with PoS, lemma, genre	33.41%	25.54%	27.03%
- with PoS and lemma	33.74%	27.86%	28.04%
BASELINE (only tokens)	31.25%	19.26%	21.33%

Table 6.2: Performance, in terms of precision (P), recall (R) and F1, of the CRF model for the event detection+classification task.

As for the task of event mention detection, all the combinations of features beat the baseline. However, information on lemma does not improve the performance of the classifier but it instead affects the precision with a difference of 0.8 points. On the contrary, PoS proved to be an important feature: without this grammatical information, all the evaluation measures significantly drop (-3.61 for precision, -5.13 for recall and -4.56 for F1 with respect to the configuration with all the features). The best feature combination, that includes PoS and genre, shows an improvement over the baseline especially in terms of recall (+8.4).

As regards event classification with no golden mentions, the results are low for all the configurations, with an F1 below 30%. Moreover, precision and recall are not balanced with a difference ranging between 4.5 and 8.28 points depending on the feature: this difference is even more evident in the baseline (11.99 points). Differently from the mention detection task, information about the lemma of each token increases both precision and recall. Eliminating PoS improves the precision (+1.85 over the configuration with all the features) but it negatively affects the recall (-1.93). The best combination of features includes only lemma and PoS with a strong improvement over the baseline in terms of recall (+8.6).

6.2. CRF CLASSIFIERS

Impact of Context Size

A second aspect to evaluate is the size of the context window around the token to be classified. To this end, we tested whether the choice of having a context window of $[+/-2]$ positions is optimal. Table 6.3 shows the performances of the CRF classifiers for event mention detection trained with the best feature selection (PoS + genre) considering different context windows: i.e., no context window (0), $[+/-1]$, $[+/-2]$, $[+/-3]$, $[+/-4]$. For each option we give the value of the macro-average precision, recall and F1. The same information is provided in Table 6.4 for the event detection+classification task.

<i>MENTION DETECTION ONLY</i>			
CONTEXT	P	R	F1
0	84.73%	69.58%	74.38%
$+/-1$	84.99%	80.13%	82.40%
$+/-2$	85.93%	83.04%	84.38%
$+/-3$	85.57%	82.70%	84.03%
$+/-4$	84.45%	81.67%	82.95%

Table 6.3: Performance of the CRF classifier on event mention detection with different context windows.

<i>DETECTION+CLASSIFICATION</i>			
CONTEXT	P	R	F1
0	41.12%	21.47%	26.11%
$+/-1$	40.16%	25.44%	28.54%
$+/-2$	33.74%	27.86%	28.04%
$+/-3$	33.69%	23.76%	25.33%
$+/-4$	32.78%	21.87%	24.74%

Table 6.4: Performance of the CRF classifier on event mention classification with different context windows.

In the detection of mention extent, the recall proves to be very sensitive to context window: by using single token features only (that is by

6.3. BI-LSTM APPROACH

considering a context window equal to 0) precision is already above 84% whereas recall is below 70%. When using a window of $[+/- 1]$ precision slightly increases (+ 0.26) but, on the contrary, recall shows an evident boost (+10.55). The best performance is achieved with a context of $[+/-2]$ that also provide balanced results between precision and recall .

Performances are less stable for the other task: in general, increasing the context makes precision worse but it also makes recall improve. More specifically, precision is higher with no context window or with a very narrow one ($[+/-1]$) while recall need a larger context. The best F1 (28.54%) is given by a context of $[+/-1]$, however precision and recall are not balanced having a difference of 14.72 points.

6.3 Bi-LSTM Approach

Our second approach is based on the use of an implementation of BiLSTM developed from the Ubiquitous Knowledge Processing Lab (Technische Universität Darmstadt)². Figure 6.2, adapted from [Reimers and Gurevych, 2017a], displays the main architecture of the system with a CRF classifier as the final layer of the network, that is with the best configuration we tested. Each word is mapped to a pre-trained word embedding and analysed to detect its casing (i.e., numeric, mainly numeric, lower case or upper case) while each character of the word is mapped to the corresponding character-level representation vector. Information about word embeddings, casing and character embeddings is concatenated to feed into the BiLSTM encoder. After the network has run from the beginning to the end of the sentence and vice versa, its output vectors are concatenated and fed to the last layer that can be a CRF classifier, as shown in the Figure, or a Softmax classifier. This second option was tested as well, together with

²<https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf>

6.3. BI-LSTM APPROACH

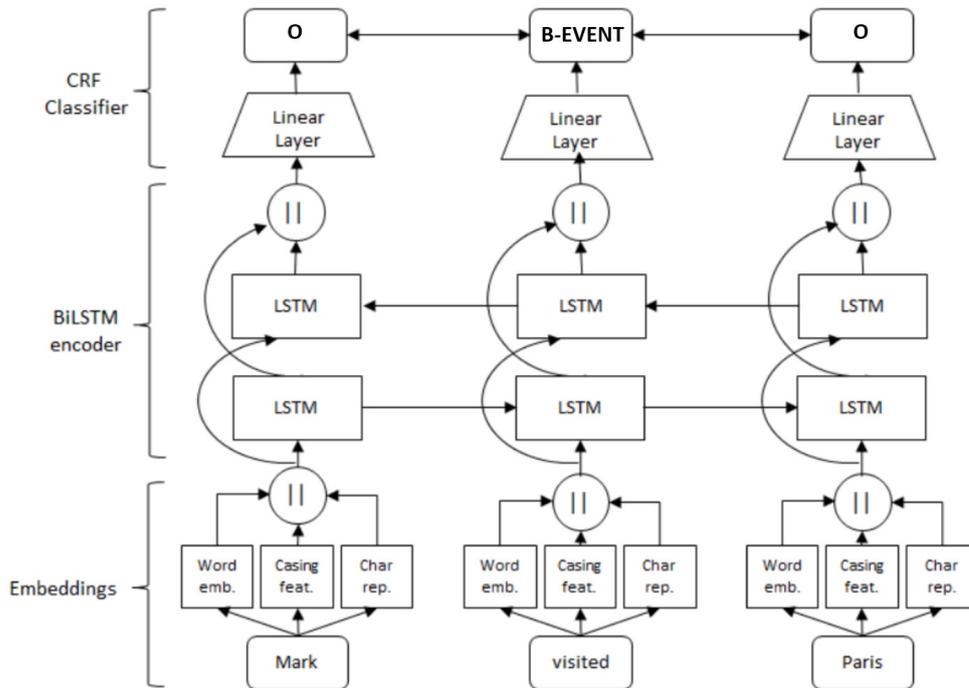


Figure 6.2: Architecture of the BiLSTM network with a CRF-classifier adapted from [Reimers and Gurevych, 2017a].

other hyperparameters, and the results are reported in the following subsection. This architecture does not require feature engineering, but only pre-trained word embeddings and a corpus of labeled data.

6.3.1 Evaluation

This subsection reports on the performances obtained on the two tasks, event mention detection only and event detection+classification, using the BiLSTM implementation previously described. Results are in terms of precision, recall and F1: given that the score of a single run is not significant because different seed values can produce very different results [Reimers and Gurevych, 2017b], we ran the system three times, took the test score from the epoch with the highest results on the development set, and then

6.3. BI-LSTM APPROACH

we calculated the average score.

As for the experimental settings, we took as a reference the setup suggested in [Reimers and Gurevych, 2017a], summarised here³:

- Mini-batch size: 8
- Recurrent units: 100
- Number of LSTM layers: 2
- Dropout: variational [0.25, 0.25]
- Classifiers: CRF
- Optimizer: nadam (Adam with Nesterov momentum) [Dozat, 2016]
- Character representation: CNNs
- Word embeddings: Komninos and Manandhar [2016]

Starting from this configuration, we performed a set of experiments changing several hyperparameters in order to identify the best options for our tasks. Below we report the results of these experiments to be compared to the ones in Table 6.5 that were obtained using the previously listed configuration.

TASK	P	R	F1
MENTION DETECTION ONLY	82.50%	83.53%	82.99%
DETECTION+CLASSIFICATION	63.46%	62.93%	63.19%

Table 6.5: Average precision (P), recall (R) and F1 over three runs of the BiLSTM system with the configuration suggested by Reimers and Gurevych [2017a].

³In their paper, Reimers and Gurevych [2017a] take into consideration various NLP tasks including event detection in accordance with the TimeML guidelines, thus considering only single-token mentions. For this task the authors test numerous configurations highlighting the hyperparameters that have an high impact on the performances in terms of F1. Our tasks are however more complex given that they include the identification of multi-token and discontinuous mentions and their classification.

6.3. BI-LSTM APPROACH

All the experiments whose results are reported in the remainder of this subsection have been carried out with an early stopping after 10 epochs if the score on the development did not increase. The implementation is based on Keras ⁴: we used Theano 1.0.0⁵ as backend.

Testing Different Hyperparameters

In the remainder of this subsection we report on the different configurations we tested: that is, six optimization algorithms, two character representations, two classifiers and nine pre-trained word embeddings.

Optimizer Optimization algorithms are used to update the model’s parameters with the aim of reducing a cost function. Over the years, different algorithms have been proposed: for example, SGD [Robbins and Monro, 1951], Adam [Kingma and Ba, 2014], Nadam [Dozat, 2016], Adagrad [Duchi et al., 2011], Adadelta [Zeiler, 2012], and RMSProp [Tieleman and Hinton, 2012]. Table 6.6 gives details on the performance of these optimizers on event mention detection while Table 6.7 contains the results of the same evaluation on event classification with no golden mentions.

In both tasks, the worst results are achieved with SGD. The difference with respect to the other optimizers is evident in particular in the classification task where SGD obtained an F1 of only 48.69% whereas all the other algorithms have an F1 above 62%. In the experiments carried out by Reimers and Gurevych [2017a], Nadam showed the best performance in all tasks: this finding is confirmed for mention detection but for event classification RMSProp achieved slightly better results, especially in terms of precision.

⁴<https://keras.io/>

⁵<http://deeplearning.net/software/theano/>

6.3. BI-LSTM APPROACH

MENTION DETECTION ONLY			
OPTIMIZER	P	R	F1
Nadam	82.5%	83.53%	82.99%
Adam	80.37%	83.47%	82.99%
SGD	79.73%	80.94%	80.51%
Adagrad	81.50%	83.30%	82.40%
Adadelta	80.15%	84.20%	82.14%
RMSProp	80.90%	83.03%	81.91%

Table 6.6: Results of the BiLSTM system with different optimization algorithms on the event mention detection only task.

DETECTION+CLASSIFICATION			
OPTIMIZER	P	R	F1
Nadam	63.46%	62.93%	63.19%
Adam	62.7%	63.90%	63.27%
SGD	51.50%	46.2%	48.69%
Adagrad	62.8%	62.43%	62.61%
Adadelta	62.97%	63.13%	63.05%
RMSProp	63.93%	62.70%	63.32%

Table 6.7: Results of the BiLSTM system with different optimization algorithms on the event detection+classification task.

Character Embeddings The architecture we adopted implements two different approaches to derive character representations: one is based on a convolution neural network (CNN) that takes into account only character trigrams without considering their position inside the word [Ma and Hovy, 2016], the other uses a BiLSTM network considering all the characters of the word and also their position, thus distinguishing between characters at the beginning, in the middle and at the end [Lample et al., 2016]. Table 6.8 shows that by using this latter approach on the mention detection task, the F1 is higher thanks to an improvement in recall (+1.7 with respect to the CNN approach). This result confirms the findings of Reimers and Gurevych [2017a], who indicate the LSTM character embeddings are the

6.3. BI-LSTM APPROACH

best performing in the TimeML event detection task.

MENTION DETECTION ONLY			
	P	R	F1
CNN	82.50%	83.53%	82.99%
LSTM	81.40%	85.23%	83.37%
NONE	82.53%	83.65%	83.04%

Table 6.8: Performance with different character embeddings options on the event mention detection only task.

As for event classification, Table 6.9 shows that the two character-based representations do not contribute much to overall performance: the difference between them is minimal with a variation of only a few decimals. In the experiments reported in [Reimers and Gurevych, 2017a], not using character embeddings is never the best option: however in our case the highest precision, recall and F1 are achieved without either of them.

DETECTION+CLASSIFICATION			
	P	R	F1
CNN	63.46%	62.93%	63.19%
LSTM	63.86%	63.30%	63.57%
NONE	63.93%	63.7%	63.81%

Table 6.9: Performance with different character embeddings options on the task of event detection+classification.

Classifier The last layer of the network can be configured as a CRF or a Softmax classifier. The main difference between the two classifiers is that in Softmax each token is seen as isolated, without considering dependencies between the tags in a sentence, whereas in CRF correlations between tags are taken into account. As reported in Tables 6.10 and 6.11, this last approach achieves better results for all three evaluation metrics in both tasks.

6.3. BI-LSTM APPROACH

MENTION DETECTION ONLY			
	P	R	F1
CRF	82.5%	83.53%	82.99%
Softmax	81.10%	82.30%	81.69%

Table 6.10: Precision, Recall and F1 score with the CRF and Softmax classifiers in the event mention detection only task.

DETECTION+CLASSIFICATION			
	P	R	F1
CRF	63.46%	62.93%	63.19%
Softmax	62.67%	62.57%	62.61%

Table 6.11: Precision, Recall and F1 score with the CRF and Softmax classifiers in the event detection+classification task.

The main issue with Softmax is that it generates invalid sequences of tags, such as B-ECONOMY I-SPACE_MOVEMENT I-ECONOMY, due to the fact that it does not maximize the tag probability of the whole sentence as CRF does. This has a negative impact on our tasks where there are strong dependencies between output tags. Our results are in contrast to the ones discussed in [Reimers and Gurevych, 2017a]: Softmax performs better in the TimeML event detection task because only single-token events are annotated, thus no information about tag dependencies is needed.

Pre-trained Embeddings In recent years, pre-trained word vectors have become important resources largely adopted to deal with many NLP tasks [Collobert et al., 2011] and many pre-trained word embeddings have been released. Beyond Komninos and Manandhar embeddings (*Komn*)⁶, we tested other resources available online, namely:

⁶<https://www.cs.york.ac.uk/nlp/extvec/>

6.3. BI-LSTM APPROACH

- GloVe, with both 300 and 100 dimensions (*GloVe300 - GloVe100*)⁷ [Pennington et al., 2014], trained on a corpus of 6 billion tokens consisting of the 2014 English Wikipedia and Gigaword 5;
- *GoogleNews*, with 300 dimensions and trained on a subset of the Google News corpus (about 100 billion words)⁸ [Mikolov et al., 2013]
- Levy and Goldberg embeddings (*Levy*)⁹, with 300 dimensions and produced from the English Wikipedia on the basis of dependency-based contexts [Levy and Goldberg, 2014]
- *fastText*, with 300 dimensions and trained on the English Wikipedia using character n-grams¹⁰ [Bojanowski et al., 2017].

By taking into consideration the previously listed pre-trained embeddings, we cover different types of word representation: GloVe and GoogleNews are based on linear bag-of-words contexts, Levy and Komn on dependency parse-trees, and fastText on a bag of character n-grams. In addition, we created historical word embeddings by processing a subset of the Corpus of Historical American English (COHA) [Davies, 2012] with GloVe, fastText and Levy and Goldberg’s code. The subset of COHA we have chosen contains 36,856 texts published between 1860 and 1939 for a total of more than 198 million words. Texts belong to four main genres (fiction, newspaper, magazine, non-fiction) balanced within each decade. The word embeddings thus trained (*HistoGlove HistoFast* and *HistoLevy*) have 300 dimensions and are publicly available online¹¹.

⁷<https://github.com/stanfordnlp/GloVe>

⁸<https://code.google.com/archive/p/word2vec/>

⁹<https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/>

¹⁰<https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md>

¹¹Our historical embeddings are available on GitHub: <https://github.com/dhfbk/Histo>. Original raw texts extracted from COHA cannot be distributed because of copyright restrictions: <https://www.corpusdata.org/restrictions.asp>

6.3. BI-LSTM APPROACH

Tables 6.12 and 6.13 contain results obtained with the tested word embeddings for event detection and classification respectively.

MENTION DETECTION ONLY			
	P	R	F1
Komn	82.5%	83.53%	82.99%
Levy	79.4%	83.13%	81.21%
GloVe300	80.07%	79.73%	79.89%
GloVe100	79.30%	80.90%	80.13%
GoogleNews	80.60%	81.70%	81.15%
FastText	79.47%	82.00%	81.25%
HistoGloVe	79.10%	82.37%	80.63%
HistoFast	79.90%	81.00%	80.44%
HistoLevy	80.70%	81.47%	81.06%

Table 6.12: Results obtained with different pre-trained word embeddings for the event mention detection only task.

DETECTION + CLASSIFICATION			
	P	R	F1
Komn	63.46%	62.93%	63.19%
Levy	62.1%	60.83%	61.44%
GloVe300	61.87%	59.57%	60.69%
GloVe100	60.23%	58.10%	59.16%
GoogleNews	63.20%	62.67%	62.93%
FastText	62.18%	61.79%	62.02%
HistoGloVe	60.00%	59.27%	59.64%
HistoFast	59.8%	55.93%	57.78%
HistoLevy	63.3%	59.40%	61.29%

Table 6.13: Results obtained with different pre-trained word embeddings for the event detection+classification task.

In both tasks, the Komninos and Manandhar embeddings perform best: the configuration including them is the only one that reaches almost 83% F1 for event detection and exceeds 63% for event classification. This means

6.3. BI-LSTM APPROACH

that capturing both semantic and syntactic similarities between words is crucial for the tasks. Also, Levy and Goldberg embeddings are dependency-based but precision with them falls below 80%. The main difference between the two representations is that Komninos and Manandhar extended the skipgram model including more types of co-occurrences within the dependency graph, thus they capture better the functional properties of words. As for GloVe, there is not much difference between the two dimensions (300 and 100): however the overall results are almost 3 points lower for event detection and 4 points lower for classification compared to the model employing the Komninos and Manandhar embeddings. No improvement is registered for either GoogleNews or fastText. As for historical embeddings, the contribution of *HistoGloVe* and *HistoFast* is not helpful, especially for precision in event detection, for which we obtained scores below 80%. The drop in performance using these embeddings is more evident in event classification than in mention detection. For the latter, F1 is about 2.4 points lower but for the former the difference ranges between 3.55 and 5.41 points. *HistoLevy*, on the contrary, performs better than *GloVe300* and *Glove100*: its F1 is slightly lower than the one achieved with the original Levy and Goldberg’s embeddings, but the precision on classification is 1.2 points higher. These scores confirm that dependency-based embeddings have a positive impact on our tasks.

Is important to note that the amount of training data has an impact on the quality of word vectors because more data produces more accurate vectors [Mikolov et al., 2013]: however, our historical word representations were trained on a corpus much smaller than the corpora used to build the other embeddings (for example, *GoogleNews* embeddings are trained on about 100 billion words, whereas the COHA subset consists of just 198.7 million words). This might be the reason that we had lower performance.

6.4 Systems Comparison and Discussion

Evaluations described in the previous Sections led us to identify the best configurations for our tasks and for the two approaches, i.e., CRF and BiLSTM.

For the task of mention detection, the best CRF classifier we release is based on a combination of three features (token, PoS and text genre) and a context window of $[\pm 2]$. For the same task, we set the neural architecture with the following parameters:

- Mini-batch size: 8
- Recurrent units: 100
- Number of LSTM layers: 2
- Dropout: variational [0.25, 0.25]
- Classifiers: CRF
- Optimizer: nadam
- Character representation: LSTM
- Word embeddings: Komninos and Manandhar [2016]

Table 6.14 reports the performances of the best models we obtained for the detection of event mentions together with the baseline, i.e. a CRF classifier trained having only tokens as features. The difference between the CRF classifier and the BiLSMT model in terms of F1 is minimal (0.76). However, it is interesting to notice that the former has a higher precision whereas the second has a higher recall. This means that the neural architecture is more able to generalize the observations of events from the training data.

6.4. SYSTEMS COMPARISON AND DISCUSSION

MENTION DETECTION ONLY			
	P	R	F1
CRF	85.93%	83.04%	84.38%
BiLSTM	82.30%	85.00%	83.62%
Baseline	80.35%	74.64%	77.14%

Table 6.14: Results of the CRF classifier and the BiLSTM model with the best configuration for the event mention detection only task.

The task dealing with both event detection and classification needed different configurations. The CRF classifier was trained with tokens, PoS and lemmas and with a context size window of $[+/-1]$. In the BiLSTM network two different hyperparameters turned out to achieve better performance with respect of the ones adopted for the mention detection only task. More specifically, we applied the RMSprop optimizer, instead of nadam, and we didn't use any character-based representation. Scores for this task are reported in Table 6.15 and compared to the baseline obtained, also in this case, by training a CRF classifier only with tokens as features.

DETECTION+CLASSIFICATION			
	P	R	F1
CRF	40.16%	25.44%	28.54%
BiLSTM	66.20%	62.70%	64.39%
Baseline	31.25%	19.26%	21.33%

Table 6.15: Results of the CRF classifier and the BiLSTM model with the best configuration for the task including both the detection and the classification of event mentions.

The neural network performs remarkably better than the CRF method with a difference of more than 26 points in terms of precision, 37 points in terms of recall and 35 as for F1. The lack of semantic and syntactic information as feature of the CRF classifier has a negative impact on the classifiers for this task. However, adding such information is not straightforward: CRFsuite supports only nominal features thus real-valued em-

6.4. SYSTEMS COMPARISON AND DISCUSSION

bedding vectors cannot be used.

Both approaches are skewed towards precision: this bias is more evident in the CRF, whereas the network produces more balanced results. A detailed comparison of the scores at the level of event classes is given in Figure 6.3.

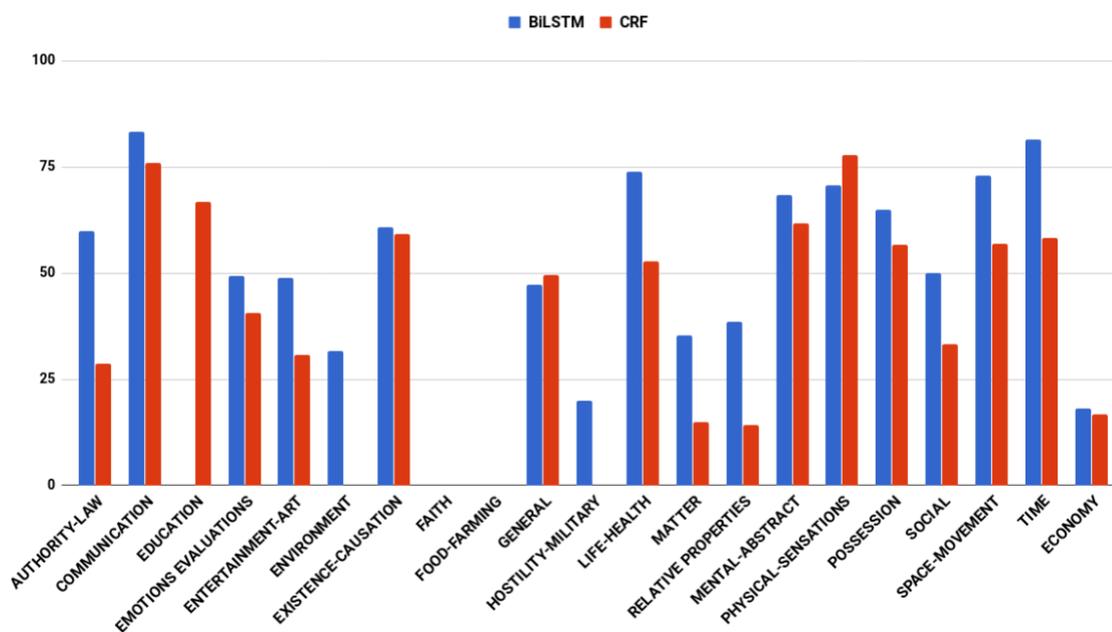


Figure 6.3: Comparison of F1 scores for each evaluated event class. The CLOTHES class is not in the Figure because it was not present in the test set.

Both approaches failed in classifying events of the classes **FOOD-FARMING** and **FAITH** that had very few occurrences in both in the training and in the test sets. The BiLSTM model wrongly classified **EDUCATION** events: in particular it assigned the class **MENTAL-ABSTRACT** to the verb “to learn”. This annotation is not totally incorrect from the semantic point of view, given that learning is a mental process. On the other side, the CRF classifier did not assign the correct value to any of the events in the **HOSTILITY-MILITARY** and **ENVIRONMENT** class. As for the latter, the score is however not high also for the BiLSTM model (F1=31.58%) because it

6.4. SYSTEMS COMPARISON AND DISCUSSION

failed in the classification of nominal events (e.g., “storm”, “tempest”) and properly classified the verb “to fall” only when the subject, belonging to the environmental domain, was close to the verb. In the following sentence, for example, “falling” is annotated with the right class whereas “fell” is annotated with the class **SPACE-MOVEMENT**: *rain commenced falling at 8:10 p.m. , and between 8:14 and 8:26 one-fifth of an inch fell.*

The different performance between the two approaches is very evident for some classes: as for **AUTHORITY-LAW**, BiLSTM is able to recognize verbs, nouns and expressions related to judicial processes and government-sanctioned practices that CRF do not even annotate as events: e.g., “to be sentenced”, “confinement”, “to be charged”. CRF also fails in the recognition of some aspectual events that BiLSTM correctly annotates with the class **TIME**: e.g. “to cease”, “to commence”.

If compared to the results of the inter-annotator agreement, we notice that three of the classes with higher F1 had also a perfect agreement between human annotators: this means that **COMMUNICATION**, **PHYSICAL SENSATIONS** and **LIFE-HEALTH** are the less ambiguous classes to be identified. On the contrary, other classes with perfect IAA have very low scores or even an F1 equal to zero: this is the case of **FOOD-FARMING**, **EDUCATION** and **ECONOMY**. The BiLSTM model, for example, correctly annotated only the verb “to pay” as belonging to the **ECONOMY** class but assigned the class **RELATIVE PROPERTIES** to copular constructions including monetary expressions, such as in “the loss **was \$ 600**”, interpreting these expressions as quantities. The same construction was not recognized as an event by the CRF classifier.

To conclude, the BiLSTM models can perform our tasks with good performance. This is particularly evident considering the task that combines both mention detection and classification for which the CRF classifier yielded much worse results.

6.5 Chapter Summary

This Chapter addresses event detection and classification from the point of view of automatic processing. First of all, we defined two tasks: one aimed at identifying the span of event mentions (*mention detection only* task) and one aimed at creating an end-to-end system that, starting from raw text, identifies event mentions and assigns them to the correct class (*detection+classification* task).

These tasks have been explored testing both traditional linear statistical models, based on CRF classifiers and hand-crafted features, and a deep learning architecture that exploits only word-based representations as features. We performed several evaluations of both the approaches: we tested different features and content size windows of the CRF classifiers and different hyperparameter options of the Bidirectional Long Short Term Memory network we adopted.

Not all the features we initially selected for the CRF proved to be helpful and the final best configuration is different, in terms of feature combination and context size, for the mention detection only task and the detection+classification task. As for the BiLSTM network, we evaluated the usage of optimization algorithms, character-based embeddings, classifiers and pre-trained word embeddings. As for this last point, we took into consideration six widely known publicly available resources and two additional word vectors we created with the GloVe and fastText method starting from a corpus of historical texts.

The final BiLSTM models we trained achieved good performances both in detecting mention extents (F1=83.62%) and in classifying events with no golden mentions given (F1=64.39%). This last result is particularly satisfactory given the complexity of the task that requires the identification of multi-token and discontinuous mentions and a classification involving

6.5. CHAPTER SUMMARY

numerous semantic classes. For the same task, the CRF classifier obtained an F1 of 28.54% only.

Chapter 7

Conclusions

In this work we have provided a theoretical and practical investigation on the topic of event detection and classification of historical texts.

After having defined our research questions in Chapter 1, in Chapter 2 we have thoroughly described how events have been defined in the field of Information Extraction by comparing evaluation campaigns and annotation guidelines devoted to the detection and processing of events. We have also addressed related topics such as the use of crowdsourcing for event annotation and the issue of multilinguality by reporting on studies we conducted during the PhD. Projects in the area of Digital Humanities have been presented in Chapter 3 in which we have also critically analysed current approaches to event detection and processing in that area. Chapter 4 contains details and results of an original case study we carried out with the aim of leveraging knowledge about the way events are defined in historical research. Chapter 5 is devoted to two other contributions of this thesis: the development of new annotation guidelines for event mention detection and classification and the release of a new manually annotated corpus of historical texts, the largest publicly available to address this task in the domain of History¹. The automatic processing of events is the topic of Chapter 6. In particular, we dealt with event mention detection and

¹The annotated corpus can be downloaded from GitHub: <https://github.com/dhfbk/Histo>.

7.1. ANSWERS TO RESEARCH QUESTIONS

classification following two approaches: traditional linear-chain CRFs and a deep learning architecture.

An innovative aspect of this thesis is given by the interdisciplinary perspective adopted in our study: more specifically, we went beyond the traditional disciplinary boundaries of NLP to integrate knowledge coming from the Humanities studies. This approach led us to develop new resources (annotation guidelines, annotated corpus and historical pre-trained word embeddings) and new models for the automatic detection and classification of event mentions².

Based on the work presented in the preceding chapters, we can now answer the research questions posed at the beginning of this thesis (Chapter 1) and provide an outlook on interesting tracks of further research.

7.1 Answers to Research Questions

Research Question 1. How can the notions of event in IE and History be combined?

Our analysis of the state of the art in both IE and DH have highlighted a lack of communication and cross-fertilization between the two research communities. This is partly due to the fact that current event definitions in IE, developed within several initiatives over the years, do not fully satisfy requirements from historians. These requirements have emerged thanks to the online questionnaire we ran that saw the direct involvement of domain experts in an ‘event definition and annotation’ exercise. The outcome of the questionnaire led us conclude that a careful adaptation of existing annotation schemes is necessary. First of all, it is important to include in the annotation durative and instantaneous happenings to-

²All the resources and the best neural models are available on GitHub: <https://github.com/dhfbk/Histo>.

7.1. ANSWERS TO RESEARCH QUESTIONS

gether with states so the broad definition of eventuality proposed by Bach [Bach, 2008], re-elaborated by Dölling [Dölling et al., 2014] and adopted in TimeML [Pustejovsky et al., 2003] better meets the needs of historians. Moreover, from the linguistic point of view, the notion of event is seen as independent from the grammatical category, so as to include not only verbs but also nouns, adjectives and other syntactic constructions. Boundaries of event mentions are not fixed thus the TimeML minimal chunk rule is not optimal, whereas the annotation of continuous and discontinuous multi-token expressions proposed in EventNugget is more in line with historians' view on events. The most relevant property of an event is its semantic type thus a semantic classification is to be preferred with respect to a categorization based on syntactic criteria. Finally, other properties characterizing events, such their level of factuality and their relations with preceding and consequent events, can be addressed by employing already available automatic modules.

Research Question 2. Can methods and techniques of Information Extraction be applied to the recognition and classification of events in historical texts in a way that it satisfies the actual needs of domain experts?

To answer this question we carried out experiments using machine learning (CRF) and deep learning (BiLSTM) techniques. These experiments were made possible by the development of new annotation guidelines for event mention detection and classification designed following historians' requirements in terms of event definition, extension, grammatical realization and classification. Our scheme has been then applied to the annotation of a corpus of historical texts including news and travel reports, a genre never addressed before in the field of Temporal Information Processing. We used the annotated corpus to test and train automatic modules for two typical

7.2. FUTURE DIRECTIONS

IE tasks: the identification of the sole event extent and the joint classification of event extent and type. The neural architecture, that exploits as features only pre-trained word vectors, achieved satisfactory results: in particular, its performance in event classification is remarkably better than those of the CRF classifier.

7.2 Future Directions

The deep neural model we developed for the task including both the detection and the classification of event mention is, to all effect, an end-to-end system that can be applied to raw texts with satisfactory results, especially for some semantic classes of events such as those related to communication, motion, mental actions and process. This system can constitute the basis for the creation of a complete framework in which to integrate other NLP systems already available to the research community. For example, modules for temporal and causal relations extraction, event factuality detection and semantic role labeling can be added on top of our system.

Our system can be also combined with other digital tools for the exploration of collections of historical texts. We can thus envisage its integration in the ALCIDE platform [Moretti et al., 2016]. ALCIDE (*Analysis of Language and Content In a Digital Environment*) is a web-based platform designed by our group to assist humanities scholars in navigating and analysing large quantities of textual data. It already contains a content processing pipeline to perform advanced linguistic search, key-concept extraction, named entities recognition and geographical analysis: our system would add an important semantic dimension to support humanities scholars in their research activity. Moreover, thanks to the availability of an intuitive graphical interface, with ALCIDE we could perform a user evaluation monitoring the actual interaction with the system by both expert

7.2. FUTURE DIRECTIONS

and lay users.

More generally, the interdisciplinary approach we adopted offered us the opportunity of posing new questions and producing new resources by looking at an IE task from a different perspective. We hope that this perspective could be adopted also for other tasks in the future so to finally fully exploit NLP techniques and methods for the processing of historical texts.

7.2. FUTURE DIRECTIONS

List of Publications

The following is the list of publications issued during the PhD.

Journal papers

- Sprugnoli, R., Tonelli, S. (2017). One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4), 485-506.
- Sprugnoli, R., Moretti, G., Bentivogli, L., Giuliani, D. (2017). Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation*, 51(2), 283-317.
- Pierpaolo, B., Malvina, N., Viviana, P., Rachele, S., Francesco, C. (2017). EVALITA Goes Social: Tasks, Data, and Community at the 2016 Edition. *Italian Journal of Computational Linguistics*, 3(1), 93-127.
- Moretti, G., Sprugnoli, R., Menini, S., Tonelli, S. (2016). ALCIDE: Extracting and visualising content from large document collections to support Humanities studies. *Knowledge-Based Systems*, 111, 100-112.
- Sprugnoli, R., Moretti, G., Tonelli, S., Menini, S. (2016). Fifty years of European history through the Lens of Computational Linguistics: the De Gasperi Project. *Italian Journal of Computational Linguistics*, 2(2), 89-100.

7.2. FUTURE DIRECTIONS

- Sprugnoli, R., Tonelli, S., Marchetti, A., Moretti, G. (2016). Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4), 762-772.
- Attardi, G., Basile, V., Bosco, C., Caselli, T., DellOrletta, F., Montemagni, S., Patti, V., Simi, M., Sprugnoli, R. (2015). State of the art language technologies for italian: The EVALITA 2014 perspective. *Intelligenza Artificiale*, 9(1), 43-61.

Book Chapters

- Caselli, T., Sprugnoli, R. (2017). It-TimeML and the Ita-TimeBank: Language Specific Adaptations for Temporal Annotation. In *Handbook of Linguistic Annotation*, pp. 969-988). Springer, Dordrecht.

Conference papers

- Rachele, S., Sara, T., Giovanni, M., Stefano, M. (2017). *A little bit of bella pianura*: Detecting Code-Mixing in Historical English Travel Writing. In *CLiC-it 2017*, pp. 304-309. Accademia University Press.
- Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., Lepri, B. (2017). RAMBLE ON: Tracing Movements of Popular Historical Figures. In *Proceedings of EACL 2017*, pp. 77-80.
- Sprugnoli, R., Caselli, T., Tonelli, S., Moretti, G. (2017). The Content Types Dataset: a New Resource to Explore Semantic and Functional Characteristics of Texts. In *Proceedings of EACL 2017*, pp. 260-266.
- Sprugnoli, R., Patti, V., Cutugno, F. (2016). Raising interest and collecting suggestions on the EVALITA evaluation campaign. In *3rd Italian Conference on Computational Linguistics, CLiC-it 2016 and*

7.2. FUTURE DIRECTIONS

5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2016, vol. 1749. CEUR-WS.

- Moretti, G., Sprugnoli, R., Tonelli, S. (2016). KD Strikes Back: from Keyphrases to Labelled Domains Using External Knowledge Sources. In Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), pp. 216-221.
- Menini, S., Sprugnoli, R., Uva, A. (2016). “Who was Pietro Badoglio?” Towards a QA system for Italian History. In Proceedings of LREC, pp. 430-435.
- Caselli, T., Sprugnoli, R., Inel, O. (2016). Temporal Information Annotation: Crowd vs. Experts. In Proceedings of LREC, pp. 3502-3509. Caselli, T., Moretti, G., Sprugnoli, R., Tonelli, S., Lanfrey, D., Kutzmann, D. S. (2016). NLP and Public Engagement: The Case of the Italian School Reform. In LREC, pp. 401-406.
- Basile, P., Cutugno, F., Nissim, M., Patti, V., Sprugnoli, R. (2016). EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language. In 3rd Italian Conference on Computational Linguistics, CLiC-it 2016 and 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2016, vol. 1749. CEUR-WS.
- Moretti, G., Sprugnoli, R., Tonelli, S. (2016). KD Strikes Back: from Keyphrases to Labelled Domains Using External Knowledge Sources. In Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), pp. 216-221.
- Minard, A. L., Speranza, M., Sprugnoli, R., Caselli, T. (2015). FacTA: Evaluation of Event Factuality and Temporal Anchoring. In Proceed-

7.2. FUTURE DIRECTIONS

ings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015), pp. 187-192.

- Moretti, G., Sprugnoli, R., Tonelli, S. (2015). Digging in the dirt: Extracting keyphrases from texts with KD. In Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015), pp. 198-203.
- Sprugnoli, R., DellOrletta, F., Caselli, T., Montemagni, S., Bosco, C. (2015). Parsing Events: a New Perspective on Old Challenges. In Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015), pp. 258-263.
- Caselli, T., Sprugnoli, R., Speranza, M., Monachini, M. (2014). EVENTI: Evaluation of Events and Temporal Information at EVALITA 2014. In Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014, pp. 27-34. Pisa University Press.
- Moretti, G., Tonelli, S., Menini, S., Sprugnoli, R. (2014). ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment. In Proceedings of the First Italian Conference on Computational Linguistics (CLiC-It 2014), pp. 270-274.
- Mirza, P., Sprugnoli, R., Tonelli, S., Speranza, M. (2014, April). Annotating causality in the TempEval-3 corpus. In Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), pp. 10-19.
- Girardi, C., Speranza, M., Sprugnoli, R., Tonelli, S. (2014, May). CROMER: a Tool for Cross-Document Event and Entity Coreference. In LREC, pp. 3204-3208.

7.2. FUTURE DIRECTIONS

- Sprugnoli, R., Lenci, A. (2014). Crowdsourcing for the identification of event nominals: an experiment. In LREC, pp. 1949-1955.

Conference abstracts

- Sprugnoli, R., Moretti, G., Tonelli, S. (2018). Temporal Dimension in Alcide De Gasperi: Past, Present and Future in Historical Political Discourse. Book of abstracts of the 7th AIUCD Conference (AIUCD 2018).
- Moretti, G., Sprugnoli, R., Tonelli, S. (2018). LETTERE: LETters Transcription Environment for REsearch. Book of abstracts of the 7th AIUCD Conference (AIUCD 2018).
- Sprugnoli, R. (2018). “Two days we have passed with the ancients...”: a Digital Resource of Historical Travel Writings on Italy. Book of abstracts of the 7th AIUCD Conference (AIUCD 2018).
- Sprugnoli, R., Caselli, T., Tonelli, S. (2017). Annotation of Content Types in Historical Texts and Contemporary News. In the book of abstract of the 27th Meeting of Computational Linguistics in the Netherlands (CLIN27).
- Sprugnoli, R., Moretti, G., Tonelli, S. (2017). Twitter Data Exploration for Italian History. In the book of abstract of the 6th AIUCD Conference (AIUCD 2017).
- Moretti, G., Tonelli, S., Sprugnoli, R. (2016). Collecting Judgments on Artworks Through a Similarity Game. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University Pedagogical University, Krakw.

7.2. FUTURE DIRECTIONS

- Elli, T., Moretti, G., Sprugnoli, R., Mauri, M., Uboldi, G., Tonelli, S., Ciuccarelli, P. (2016). Visualisation Strategies for Comparing Political Ideas with the ORATIO Platform. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University Pedagogical University, Krakw.
- Caselli, T., Sprugnoli, R. (2015). Crowdsourcing Temporal Relations in Italian and English. In the book of abstract of the 25th Meeting of Computational Linguistics in the Netherlands (CLIN25).
- Marchetti, A., Sprugnoli, R., Tonelli, S.. (2014). Sentiment Analysis for the Humanities: the Case of Historical Texts. In Digital Humanities 2014: Conference Abstracts.

In press

- Sprugnoli, R., Moretti, G., Tonelli, S. (IN PRESS). LOD Navigator: Tracing Movements of Italian Shoah Victims. To appear in Proceedings of EHRI workshop Data Sharing, Holocaust Documentation, Digital Humanities.
- Speranza, M., Sprugnoli, R. (IN PRESS). Annotation of Temporal Information on Historical Texts: a Small Corpus for a Big Challenge. To appear in Proceedings of the "Formal Representation and Digital Humanities" workshop.

Bibliography

Jean-Michel Adam. Quels types de textes?(What Kinds of Text?). *Français dans le monde*, 192:39–43, 1985.

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. A Comparison of the Events and Relations Across ACE, ERE, TAC-KBP, and FrameNet Annotation Standards. In *Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W14-2907>.

James F Allen. Towards a General Theory of Action and Time. *Readings in Planning*, 23(2):464–479, 1990. ISSN 00043702. doi: 10.1016/0004-3702(84)90008-0.

James F Allen and George Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579, 1994.

Jeffrey Allen and Khalid Choukri. Survey of Language Engineering Needs: a Language Resources Perspective. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, 2000.

Pascal Amsili, Pascal Denis, and Laurence Danlos. French TimeBank : An ISO-TimeML Annotated Reference Corpus. In *ACL short*, pages

BIBLIOGRAPHY

130–134, Portland, Oregon, USA, 2011. Association for Computational Linguistics. ISBN 9781932432886.

Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. Sra: Description of the ie2 system used for muc-7. In *Proceedings of the seventh message understanding conference (MUC-7)*, pages 123–135, 1998.

Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, and Mabry Tyson. FASTUS: A finite-state processor for information extraction from real-world text. In *International Joint Conference on Artificial Intelligence*, volume 13, pages 1172–1178, vol. 93, . Chambéry, France, 1993. International Joint Conference on Artificial Intelligence.

Ju D Apresjan. Regular polysemy. *Linguistics*, 12(142):5–32, 1974.

Dawn Archer. Exploring verbal aggression in English historical texts using USAS: the possibilities, the problems and potential solutions. In I Taavitsainen, A H Jucker, and J Tuominen, editors, *Diachronic corpus pragmatics*, pages 277–302. John Benjamins Publishing Company, University of Helsinki / University of Zurich, 2014. doi: 10.1075/pbns.243.17arc.

Lora Aroyo and Chris Welty. Harnessing Disagreement for Event Semantics. In *Proceedings of DeRiVE 2012 Workshop, ISWC*, 2012.

Lora Aroyo and Chris Welty. The three sides of CrowdTruth. *Journal of Human Computation*, 1:31–34, 2014.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Masayuki Asahara. BCCWJ-TimeBank : Temporal and Event Information Annotation on Japanese Text. In *Pacific*, volume 19, pages 1–24,

BIBLIOGRAPHY

- Taipei, Taiwan, 2013. Association for Computational Linguistics. ISBN 9789860385670.
- Naveen Ashish, Doug Appelt, Dayne Freitag, and Dmitry Zelenko. Proceedings of the Workshop on Event Extraction and Synthesis. Technical report, WS-06-07, American Association for Artificial Intelligence, 2006.
- Farzindar Atefeh and Wael Khreich. a Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31(1):n/a–n/a, 2013. ISSN 1467-8640. doi: 10.1111/coin.12017. URL <http://doi.wiley.com/10.1111/coin.12017>.
- Emmon Bach. The Algebra of Events. *Formal Semantics: The Essential Readings*, 9(1):324–333, 2008. ISSN 01650157. doi: 10.1002/9780470758335.ch13.
- Daniela Baiamonte, Tommaso Caselli, and Irina Prodanof. Annotating Content Zones in News Articles. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale, 2016.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics -*, volume 1, page 86, Montréal, Quebec, Canada, 1998. Association for Computational Linguistics. ISBN 978-88-07-72177-9. doi: 10.3115/980845.980860. URL <http://portal.acm.org/citation.cfm?doid=980845.980860>.
- Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. CAT: the CELCT Annotation Tool. In *Proceedings of LREC 2012*, pages 333–338, 2012.

BIBLIOGRAPHY

Waitman Beorn, Tim Cole, Simone Gigliotti, Alberto Giordano, Anna Holian, Paul B Jaskot, Anne Kelly Knowles, Marc Masurovsky, and Erik B Steiner. Geographies of the Holocaust. *Geographical Review*, 99(4):563–574, 2009.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado, USA, 2015a. Association for Computational Linguistics. doi: 10.18653/v1/S15-2136. URL <http://aclweb.org/anthology/S15-2136>.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, 2015b.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California, 2016. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S16-1165>.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada, 2017. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S17-2093>.

Douglas Biber. A typology of English texts. *Linguistics*, 27(1):3–44, 1989.

BIBLIOGRAPHY

- Heike Bieler, Stefanie Dipper, and Manfred Stede. Identifying formal and functional zones in film reviews. *Proceedings of the 8th SIGDIAL*, pages 75–78, 2007.
- Jari Björne and Tapio Salakoski. Generalizing biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 183–191. Association for Computational Linguistics, 2011.
- Jari Björne and Tapio Salakoski. TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task. In *BioNLP Shared Task 2013 Workshop*, pages 16–25, Sofia, Bulgaria, 2013. Association for Computational Linguistics.
- Marc Bloch. *The Historian’s Craft: Translated from the French by Peter Putnam*. Manchester University Press, 1954.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- Laura Brazzo and Silvia Mazzini. From the Holocaust Victims Names to the Description of the Persecution of the European Jews in Nazi Years: the Linked Data Approach and a New Domain Ontology. In *Book of abstract of DH 2015*, 2015.
- Laura Brazzo and Silvia Mazzini. Linked Open Data per l’analisi dei dati e lo sviluppo della ricerca sulla Shoah in Italia. In *Book of abstract of AIUCD 2017*, 2017.
- Michael Buckland and Michele Renee Ramos. Events as a structuring device in biographical mark-up and metadata. *Bulletin of the Association for Information Science and Technology*, 36(2):26–29, 2010.
- Peter Burke. *Varieties of cultural history*. Cornell University Press, 1997.

BIBLIOGRAPHY

Gully A P C Burns, Pradeep Dasigi, Anita de Waard, and Eduard H Hovy. Automated detection of discourse segment and experimental types from the text of cancer pathway results sections. *Database*, 2016:baw122, 2016.

James Buzard. *The Grand Tour and after (1660-1840)*, page 3752. Cambridge Companions to Literature. Cambridge University Press, 2002. doi: 10.1017/CCOL052178140X.003.

Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

Tommaso Caselli and Irina Prodanof. Robust Temporal Processing: from Model to System. *Special issue: Natural Language Processing and its Applications*, 46:29–40, 2010.

Tommaso Caselli and Rachele Sprugnoli. It-TimeML: TimeML Annotation Guidelines for Italian. Technical report, Technical Report, 2015. URL <https://sites.google.com/site/ittimeml/documents>.

Tommaso Caselli and Rachele Sprugnoli. It-TimeML and the Ita-TimeBank: Language Specific Adaptations for Temporal Annotation. In *Handbook of Linguistic Annotation*, pages 969–988. Springer, 2017.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. Annotating Events, Temporal Expressions and Relations in Italian : the It-TimeML Experience for the Ita-TimeBank. In *Proceedings of the Fifth Law Workshop (LAW V)*, number June, pages 23–24, 2011a. ISBN 9781932432930.

Tommaso Caselli, Hector Llorens, Borja Navarro-Colorado, and Estela Saquete. Data-driven approach using semantics for recognizing and classifying TimeML events in Italian. In *Proceedings of the International*

BIBLIOGRAPHY

- Conference Recent Advances in Natural Language Processing 2011*, pages 533–538, 2011b.
- Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 27–34. Pisa University Press, 2014a.
- Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. EVENTI EValuation of Events and Temporal INformation at Evalita 2014. In Italy Pisa, editor, *Proceedings of the Fourth International Workshop EVALITA 2014*. Pisa University Press, 2014b.
- Tommaso Caselli, Rachele Sprugnoli, and Oana Inel. Temporal information annotation: Crowd vs. experts. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016. ISBN 9782951740891.
- Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. In *LREC*, 2016.
- Seymour Benjamin Chatman. *Coming to terms: the rhetoric of narrative in fiction and film*. Cornell University Press, 1990.
- Nancy A. Chinchor. Overview of MUC-7/MET-2. In *Message understanding conference (muc-7) proceedings*, VA, 1998. Fairfax. URL ieeexplore.ieee.org/document/7050708/.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost)

BIBLIOGRAPHY

from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Francisco Costa and António Branco. Timebankpt: A timeml annotated corpus of portuguese. In *LREC*, pages 3727–3734, 2012.

Dallas Costis, Nephelie Chatzidiakou, Agiatis Benardou, Claire Clivaz, John Cunningham, Meredith Dabek, Patricia Garrido, Elena Gonzalez-Blanco, Jurij Hadalin, Lorna Hughes, et al. European survey on scholarly practices and digital needs in the arts and humanities digital methods and practices observatory working group (dimpo) survey highlights en. Technical report, DARIAH; DiMPO Working Group, 2017.

Agata Cybulska and Piek Vossen. Event Models for Historical Perspectives: Determining Relations between High and Low Level Events in Text, Based on the Classification of Time, Location and Participants. In *Proceedings of LREC*, 2010.

Agata Cybulska and Piek Vossen. Historical event extraction from text. In *Proceedings of the 5th ACL-HLT Workshop on ...*, number June, pages 39–43. Association for Computational Linguistics, 2011. ISBN 9781937284046. URL <http://dl.acm.org/citation.cfm?id=2107642>.

Agata Cybulska and Piek Vossen. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In Iceland Reykjavik, editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552. European Language Resources Association (ELRA, 2014. ISBN 978-2-9517408-8-4.

Ulrike Czeitschner and Barbara Krautgartner. Discursive Constructions of Culture: Semantic Modelling for Historical Travel Guides. *Sociology and Anthropology*, 5(4):323–331, 2017.

BIBLIOGRAPHY

- Vahakn N Dadrian. Patterns of twentieth century genocides: the Armenian, Jewish, and Rwandan cases. *Journal of Genocide Research*, 6(4): 487–522, 2004.
- Mariana Damova and Sabine Bergler. The aspectual type begin. In *Conference TALN*, 2000.
- Donald Davidson. *Essays on actions and events: Philosophical essays*, volume 1. Oxford University Press on Demand, 2001.
- Mark Davies. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157, 2012.
- Victor De Boer, Johan Oomen, Oana Inel, Lora Aroyo, Elco Van Staveren, Werner Helmich, and Dennis De Beurs. DIVE into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:152–158, 2015.
- Alcide De Gasperi. Alcide De Gasperi nel Trentino asburgico. In *Scritti e discorsi politici di Alcide De Gasperi*, volume 1. Il Mulino, 2006.
- Tullio De Mauro. Dizionario italiano. *Torino: Paravia*, 2000.
- Anita De Waard and Henk Pander Maat. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55. Association for Computational Linguistics, 2012.
- Marcel den Dikken and Teresa O’Neill. Copular Constructions in Syntax, 2016. URL <http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-137>.

BIBLIOGRAPHY

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In Portugal Lisbon, editor, *Proceedings of LREC 2004*. European Language Resources Association (ELRA), 2004.

Martin Doerr. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75, 2003.

Johannes Dölling et al. Aspectual coercion and eventuality structure. *Events, arguments, and aspects. Topics in the semantics of verbs*, pages 189–226, 2014.

Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Amosse Edouard, Elena Cabrio, Sara Tonelli, and Nhan Le Thanh. Graph-based Event Extraction from Twitter. In *Proceedings of Recent Advances in Natural Language Processing (RANLP17)*, 2017.

Marieke Van Erp, Anneleen Schoen, and Chantal Van Son. MEANTIME , the NewsReader Multilingual Event and Time Corpus. In *Proceedings of LREC 2016*, pages 4417–4422, Portorož, Slovenia, 2008. European Language Resources Association (ELRA).

BIBLIOGRAPHY

- Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.
- Liliana Picciotto Fargion. *Il libro della memoria: gli Ebrei deportati dall’Italia (1943-1945)*. Mursia, 1991.
- Lucien Febvre. *Combats pour l’histoire*. Armand Colin, Paris, 1953. ISBN 2-266-06911-X.
- Christiane Fellbaum. *WordNet*. MIT Press, Cambridge, 1998.
- Elena Filatova and Vasileios Hatzivassiloglou. Domain-Independent Detection, Extraction, and Labeling of Atomic Events. In Bulgaria Borovetz, editor, *Proceedings of RANLP*, pages 145–152, 2003.
- Elena Filatova and Eduard Hovy. Assigning time-stamps to event-clauses. In *Proceedings of the ACL-EACL 2001 Workshop for Temporal and Spatial Information Processing*, Toulouse, France, 2001. Association for Computational Linguistics.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.
- FLaReNet Working Group. Results of the questionnaire on the priorities in the field of Language Resources. Technical report, Department of Computer Science, Michigan State University, 2010.
- Antske Fokkens, Serge Ter Braake, Niels Ockeloën, Piek Vossen, Susan Legêne, and Guus Schreiber. BiographyNet: Methodological Issues when

BIBLIOGRAPHY

- NLP supports historical research. In *Proceedings of LREC*, pages 3728–3735, 2014.
- Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. Biographynet: Extracting relations between people and events. In *Biographical Data in a Digital World 2017*, 2018.
- Antske Fokkens-Zwirello, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. GAF: A grounded annotation framework for events. In *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 11–20. Association for Computational Linguistics, 2013.
- Corina Forascu and Dan Tufi. Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- Johanna Fulda, Matthew Brehmel, and Tamara Munzner. Timelinecurator: Interactive authoring of visual timelines from unstructured text. *IEEE transactions on visualization and computer graphics*, 22(1):300–309, 2016.
- Robert Gaizauskas, Kevin Humphreys, Hamish Cunningham, and Yorick Wilks. University of Sheffield: description of the LaSIE system as used for MUC-6. In *Proceedings of the 6th conference on Message understanding*, pages 207–220, VA, 1995. Fairfax. ISBN 1558604022.
- Fredric Gey, Ryan Shaw, Ray Larson, Michael Buckland, Barry Pateman, and Dan Melia. Marking Up Cultural Materials for Time and Geogra-

BIBLIOGRAPHY

- phy. In *Proceedings of the Workshop on Information Access to Cultural Heritage, Aarhus, Denmark, Sept*, volume 28, 2008.
- Paolo Giaccaria and Claudio Minca. *Hitler's Geographies: The Spatialities of the Third Reich*. University of Chicago Press, 2016.
- Alberto Giordano and Anna Holian. Retracing the hunt for jews: A spatio-temporal analysis of arrests during the holocaust in Italy. In *Geographies of the Holocaust*. Indiana University Press, 2014.
- Ralph Grishman. The Impact of Task and Corpus on Event Extraction Systems. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta., 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL <http://aclweb.org/anthology/L10-1389>.
- Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an Ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.
- Jo Guldi and David Armitage. *The History Manifesto*. Cambridge University Press, Cambridge, 2014. ISBN 9781107076341. doi: 10.1017/9781139923880. URL <https://www.cambridge.org/core/books/the-history-manifesto/AC1A1EC711AE91A4F9004E7582D79AFD>.
- Claude Hagège. *L'Homme de paroles: Contribution linguistique aux sciences humaines*, volume 13. Fayard, Paris, 1996. ISBN 2213674116.
- Kai Hakala and Sofie Van Landeghem. EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. In *Acl 2013*, pages 26–34, Sofia, Bulgaria, 2013. Association for Computational Linguistics.
- James Higginbotham. *On events in linguistic semantics*, volume 49. Oxford University Press Oxford, 2000.

BIBLIOGRAPHY

Raul Hilberg. *The destruction of the European Jews*, volume 3. Holmes & Meier New York, 1985.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. Events are Not Simple: Identity, Non-Identity, and Quasi-Identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, number June, pages 21–28, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W13-1203>.

Eero Hyvönen, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu Mäkelä. Publishing Second World War History as Linked Data Events on the Semantic Web. *Abstracts of Digital Humanities 2016, short papers*, pages 571–573.

Eero Hyvönen, Miika Alonen, Esko Ikkala, and Eetu Mäkelä. Life stories as event-based linked data: case semantic national biography. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pages 1–4. CEUR-WS. org, 2014.

Eero Hyvönen, Jouni Tuominen, Eetu Mäkelä, Jérémie Dutruit, Kasper Apajalahti, Erkki Heino, Petri Leskinen, and Esko Ikkala. Second World War on the Semantic Web: The WarSampo Project and Semantic Portal. In *International Semantic Web Conference (Posters & Demos)*, 2015.

Eero Hyvönen, Petri Leskinen, Erkki Heino, Jouni Tuominen, and Laura Sirola. Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web. In *International Conference on Language, Data and Knowledge*, pages 113–119. Springer, 2017.

Nancy Ide and Catherine Macleod. The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Lin-*

BIBLIOGRAPHY

- guistics*, volume 3, pages 1–7. Lancaster University Centre for Computer Corpus Research on Language Lancaster, UK, 2001.
- Nancy Ide and David Woolner. Exploiting Semantic Web Technologies for Intelligent Access to Historical Documents. In *Computer*, pages 2177–2180, Reykjavik, Iceland, 2004. European Language Resources Association (ELRA).
- Nancy Ide and David Woolner. Historical ontologies. *Words and Intelligence II*, pages 137–152, 2007.
- Rei Ikuta, William F Styler Iv, Mariah Hamang, Tim O Gorman, Martha Palmer, Will Styler, and Tim O’Gorman. Challenges of Adding Causation to Richer Event Descriptions. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 12–20, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W/W14/W14-2903>.
- Seohyun Im, Hyunjo You, Hayun Jang, Seungho Nam, and Hyopil Shin. KTimeML: specification of temporal and event expressions in Korean text. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 115–122, Suntec, Singapore, 2009. Association for Computational Linguistics.
- Oana Inel, Lora Aroyo, Chris Welty, and Robert-Jan Sips. Exploiting Crowdsourcing Disagreement with Various Domain-Independent Quality Measures. In *Proceedings of DeRiVE 2013 Workshop, ISWC*, 2013.
- Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. CrowdTruth: machine-human computation framework for har-

BIBLIOGRAPHY

nessing disagreement in gathering annotated data. In *The Semantic Web-ISWC 2014*, pages 486–504. Springer, 2014.

Ander Intxaurreondo, Eneko Agirre, Oier Lopez De Lacalle, and Mihai Surdeanu. Diamonds in the rough: Event extraction from imperfect microblog data. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 641–650, 2015.

Iso, SemAf/Time Working Group. ISO DIS 24617-1: 2008 Language resource management - Semantic annotation framework - Part 1: Time and events. Technical report, ISO Central Secretariat, Geneva, 2008.

Elisabetta Jezek. Polysemy of Italian event nominals. *Faits des langues*, 30:251–264, 2008.

Hyuckchul Jung and Amanda Stent. ATT1: Temporal Annotation Using Big Windows and Rich Syntactic and Semantic Features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 20–24, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

Hyuckchul Jung, James Allen, Nate Blaylock, Will De Beaumont, Lucian Galescu, and Mary Swift. Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 workshop*, pages 146–154. Association for Computational Linguistics, 2011.

Graham Katz and Fabrizio Arosio. The annotation of temporal information in natural language sentences. In *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, volume 13, page 15,

BIBLIOGRAPHY

- Toulouse, France, 2001. Association for Computational Linguistics. doi: 10.3115/1118238.1118252.
- Christian Kay, Jane Roberts, Michael Samuels, and Irené Wotherspoon. *Historical Thesaurus of the Oxford English Dictionary*. Oxford University Press, 2009a.
- Christian Kay, Jane Roberts, Michael Samuels, and Irene Wotherspoon. 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 University Press, Oxford, 2009b.
- Suzanne Kemmer and Arie Verhagen. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics*, 5(2):115–156, 1994.
- Houda Khrouf and Raphaël Troncy. EventMedia: A LOD dataset of events illustrated with media. *Semantic Web*, 7(2):193–199, 2016.
- Jin-Dong Kim, Tomoko Ohta, Tomoko Tateisi, and Junichi Tsujii. GENIA corpus manual. Technical report, 2006.
- Jin Dong Kim, Tomoko Ohta, and Jun’ichi Tsujii. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10), 2008. ISSN 14712105. doi: 10.1186/1471-2105-9-10.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Lauren F Klein and Matthew Gold. *Debates in the Digital Humanities 2016*. University of Minnesota Press, 2016.
- Anne Kelly Knowles, Tim Cole, and Alberto Giordano. *Geographies of the Holocaust*. Indiana University Press, 2014.

BIBLIOGRAPHY

Anne Kelly Knowles, Levi Westerveld, and Laura Strom. Inductive visualization: A humanistic alternative to GIS. *GeoHumanities*, 1(2):233–265, 2015.

Oleksandr Kolomiyets and Marie-Francine Moens. KUL: A data-driven approach to temporal parsing of documents. In *Proceedings of the second joint conference on lexical and computational semantics (* SEM)*, volume 2, pages 83–87, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics. URL <https://lirias.kuleuven.be/handle/123456789/401831>.

Alexandros Komninos and Suresh Manandhar. Dependency based embeddings for sentence classification tasks. In *HLT-NAACL*, pages 1490–1500, 2016.

Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, and Others. Crowdsourcing in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 10(3):177–243, 2016.

Hans-Ulrich Krieger and Thierry Declerck. TMOThe Federated Ontology of the TrendMiner Project. In *Proceedings of LREC*, pages 4164–4171, 2014.

Hans-Ulrich Krieger and Thierry Declerck. An OWL Ontology for Biographical Knowledge. Representing Time-Dependent Factual Knowledge. In *Proceedings of the First Conference on Biographical Data in a Digital World 2015 (BD 2015)*, pages 101–110, 2015.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016.

BIBLIOGRAPHY

- Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/N16-1030>.
- J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- Egoitz Laparra, Itziar Aldabe, and German Rigau. Document level time-anchoring for timeline extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 358–364, 2015.
- Patrick Le Boeuf, Martin Doerr, Christian Emil Ore, and Stephen Stead. Definition of the CIDOC conceptual reference model, Version 6.2.2. Technical report, ICOM/CIDOC Documentation Standards Group. CIDOC CRM Special Interest Group, 2017.
- Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. UHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1292–1297, San Diego, California, 2016. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S16-1201>.
- Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. Joint Entity and Event Coreference Resolution across Documents. In *(EMNLP-CoNLL 2012) Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, number July, pages 489–500, Jeju, South Korea, 2012. Association for Computational Linguistics. ISBN 9781937284435. URL <http://www.aclweb.org/anthology/D12-1045>.

BIBLIOGRAPHY

- Jake Lever and Steven J M Jones. VERSE: Event and relation extraction in the BioNLP 2016 Shared Task. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 42–49. Association for Computational Linguistics, 2016.
- Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *ACL (2)*, pages 302–308, 2014.
- Chen Li, Zhiqiang Rao, and Xiangrong Zhang. LitWay, Discriminative Extraction for Different Bio-Events. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 32–41. Association for Computational Linguistics, 2016.
- Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000, 2012.
- Linguistic Data Consortium. ACE (Automatic Content Extraction) English annotation guidelines for events. Technical Report version 5.4.3 2005.07.01, LDC, 2005. URL <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf>.
- Hector Llorens, Estela Saquete, and Borja Navarro. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *ACL 2010 - SemEval 2010 - 5th International Workshop on Semantic Evaluation, Proceedings*, pages 284–291, Uppsala, Sweden, 2010. Association for Computational Linguistics. ISBN 1932432701.
- Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. SemEval-2015 Task

BIBLIOGRAPHY

- 5: QA TempEval - Evaluating Temporal Information Understanding with Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, number SemEval, pages 46–54, Denver, Colorado, USA, 2015. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S15-2005>.
- Robert E Longacre. *The grammar of discourse*. Springer Science & Business Media, 2013.
- Peter Lunenfeld, Anne Burdick, Johanna Drucker, Todd Presner, and Jeffrey Schnapp. *Digital humanities*. MIT Press, Cambridge, MA, 2012. ISBN 9780262018470 0262018470.
- Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/P16-1101>.
- Eetu Mäkelä. Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In *European Semantic Web Conference*, pages 424–428. Springer, 2014.
- Inderjeet Mani and Barry Schiffman. Temporally anchoring and ordering events in news. *Time and Event Recognition in Natural Language.*, 2005.
- Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

BIBLIOGRAPHY

- Henri-Irénée Marrou. *De la connaissance historique*. 1954a. ISBN 2020043017 9782020043014.
- Henri-Irénée Marrou. *De la connaissance historique*. éditions du Seuil Paris, 1954b. ISBN 2020043017 9782020043014.
- Michele Mauri, Tommaso Elli, Giorgio Caviglia, Giorgio Ubaldi, and Matteo Azzi. Rawgraphs: A visualisation platform to create open outputs. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, page 28. ACM, 2017.
- Louise McNally. Existential sentences with existential quantification. *Linguistics and Philosophy*, 21(4):353–392, 1998.
- Carlo Meghini, Roberto Scopigno, Julian Richards, Holly Wright, Guntram Geser, Sebastian Cuy, Johan Fihn, Bruno Fanini, Hella Hollander, Franco Niccolucci, and Others. ARIADNE: A Research Infrastructure for Archaeology. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(3):18, 2017.
- Farrokh Mehryary, Jari Björne, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. Deep learning with minimal training data: Turkunlp entry in the bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 73–81. Association for Computational Linguistics, 2016.
- Chiara Melloni and Others. Nominals, polysemy, and co-predication. *Journal of cognitive science*, 12(1):1–31, 2011.
- Stefano Menini, Rachele Sprugnoli, Giovanni Moretti, Enrico Bignotti, Sara Tonelli, and Bruno Lepri. RAMBLE ON: Tracing movements of popular historical figures. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of the Software Demonstrations*, 2017. ISBN 9781510838604.

BIBLIOGRAPHY

- Albert Meroño-Peñuela, Ashkpour Ashkpour, Marieke van Erp, Keez Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank van Harmelen. Semantic technologies for historical research: A survey. *Semantic Web Journal*, 6(6):539–564, 2015.
- Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North ...*, pages 646–655, Montreal, Canada, 2012. Association for Computational Linguistics. ISBN 978-1-937284-20-6. URL <http://www.aclweb.org/anthology/N12-1083>.
- Line Mikkelsen. *Copular clauses: Specification, predication and equation*, volume 85. John Benjamins Publishing, 2005.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. Description of the BBN System Used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, 1998.
- Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, Rubén Urizar, and Fondazione Bruno Kessler. SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, Colorado, USA, 2015. Association for Computational Linguistics. URL <http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval132.pdf>.

BIBLIOGRAPHY

- Louis O Mink. Narrative form as a cognitive instrument. *The writing of history: Literary form and historical understanding*, pages 129–149, 1978.
- Paramita Mirza and Sara Tonelli. An Analysis of Causality between Events and its Relation to Temporal Information. In *Coling*, pages 2097–2106, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics. ISBN 9781941643266.
- Paramita Mirza and Sara Tonelli. CATENA: CAusal and TEmporal relation extraction from NATural language texts. In *Proceedings of COLING 2016 Technical Papers.*, Osaka, Japan, 2016. The 26th International Conference on Computational Linguistics.
- Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, 2014.
- Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. Event Nugget Annotation: Processes and Issues. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 66–76, Denver, Colorado, USA, 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0809.
- Monica Monachini, Anika Nicolosi, and Alberto Stefanini. Digital classics: A survey on the needs of ancient greek scholars in italy. In *Proceedings of the CLARIN Annual Conference*, 2017.
- Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. ALCIDE: Extracting and visualising content from large document collections to support humanities studies. *Knowledge-Based Systems*,

BIBLIOGRAPHY

- 111:100–112, nov 2016. ISSN 09507051. doi: 10.1016/j.knosys.2016.08.003. URL <http://linkinghub.elsevier.com/retrieve/pii/S0950705116302635>.
- Andrea Moro. *The raising of predicates: Predicative noun phrases and the theory of clause structure*, volume 80. Cambridge University Press, 1997.
- Justin Mott, Zhiyi Song, Ann Bies, and Stephanie Strassel. Parallel Chinese - English Entities, Relations and Events Corpora. In *Lrec 2016*, pages 3717–3722, Portorož, Slovenia., 2016. European Language Resources Association (ELRA). ISBN 9782951740891.
- Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria, 2013. Association for Computational Linguistics.
- Jun-Ping Ng and Min-Yen Kan. Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations. In *Proceedings of COLING 2012*, pages 2109–2124. The COLING 2012 Organizing Committee, 2012. URL <http://www.aclweb.org/anthology/C12-1129>.
- Guerrero Nieto, Saurí M., and Bernabé Poveda M A Roser. Procesamiento del lenguaje natural. *ModeS TimeBank: A Modern Spanish TimeBank Corpus*, 47:259–267, 2011.
- Melissa Terras Nyhan Julianne and Edward Vanhoutte, editors. *Defining Digital Humanities*. Ashgate, 2013.
- Michael Oakeshott. *Experience and its Modes*. Cambridge University Press, 2015.

BIBLIOGRAPHY

- Niels Ockeloën, Antske Fokkens, Serge Ter Braake, Piek Vossen, Victor De Boer, Guus Schreiber, and Susan Legêne. Biographynet: Managing provenance at multiple levels and from different perspectives. In *Proceedings of the 3rd International Conference on Linked Science-Volume 1116*, pages 59–71. CEUR-WS. org, 2013.
- Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007. URL <http://www.chokkan.org/software/crfsuite/>.
- Nelleke Oostdijk and Lou Boves. User requirements analysis for the design of a reference corpus of written Dutch. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation(LREC 2006)*, 2006.
- Sharon Ouditt and Loredana Polezzi. Introduction: Italy as place and space. *Studies in Travel Writing*, 16(2):97–105, 2012.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005. ISSN 0891-2017. doi: 10.1162/0891201053630264. URL <http://www.mitpressjournals.org/doi/10.1162/0891201053630264>.
- Alessio Palmero Aproso and Claudio Giuliano. The Wiki Machine: an open source software for entity linking and enrichment. In *ArXiv e-prints*, 2016.
- Terence Parsons. *Events in the Semantics of English*, volume 5. Cambridge, Ma: MIT Press, 1990.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 con-*

BIBLIOGRAPHY

- ference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- K.R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge & K. Paul, 1963. URL https://books.google.it/books?id=_PXWAAAAMAAJ.
- J Pustejovsky. The syntax of event structure. *Cognition*, 41(1-3):47–81, 1991a.
- James Pustejovsky. The generative lexicon. *Computational linguistics*, 17(4):409–441, 1991b.
- James Pustejovsky. A Survey of Dot Objects. Manuscript 2, 2005.
- James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* ” O’Reilly Media, Inc.”, 2012.
- James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Beth Sundheim, David Day, Lisa Ferro, and Dragomir. The TIMEBANK Corpus. In *Natural Language Processing and Information Systems*, volume 4592, pages 647–656, Lancaster, UK, 2002. UCREL. ISBN 978-3-540-73350-8. doi: 10.1007/978-3-540-73351-5. URL <http://www.springerlink.com/content/c9313110264107m6>.
- James Pustejovsky, José Castano, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of IWCS-5*, Tilburg, The Netherlands, 2003. American Association for Artificial Intelligence.

BIBLIOGRAPHY

- James Pustejovsky, Jessica Littman, and Roser Saurí. Arguments in TimeML: Events and Entities. In F Schilder, G Katz, and J Pustejovsky, editors, *Annotating, Extracting and Reasoning about Time and Events (Lecture Notes in Artificial Intelligence No. 4795)*, pages 107–126. Springer Berlin Heidelberg, Berlin, 2007. ISBN 9783540759881. doi: 10.1007/978-3-540-75989-8_8.
- Michele R. Ramos. Biography Light Ontology: An Open Vocabulary For Encoding Biographic Texts. Technical report, Bringing Lives to Light: Biography in Context Project, 2009. URL <http://metadata.berkeley.edu/BiographyLightOntology.pdf>.
- Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. The UCREL semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*, pages 7–12, Lisbon, Portugal, 2004. URL <http://eprints.lancs.ac.uk/1783/>.
- Paul Rayson, Alistar Baron, Scott Piao, and Steve Wattam. Large-scale time-sensitive semantic analysis of historical corpora. In *Proceedings of the 36th Meeting of ICAME*, Trier, Germany, 2015. ICAME.
- Nils Reimers and Iryna Gurevych. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*, 2017a.
- Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, 2017b.
- Alan Ritter, Oren Etzioni, and Sam Clark. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD international*

BIBLIOGRAPHY

- conference on Knowledge discovery and data mining*, pages 1104–1112, Beijing, China. ACM.
- Livio Robaldo, Tommaso Caselli, Irene Russo, and Matteo Grella. From Italian Text to TimeML Document via Dependency Parsing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 177–187. Springer Berlin / Heidelberg, 2011.
- Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- Pierre Sablayrolles. The semantics of motion. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 281–283. Morgan Kaufmann Publishers Inc., 1995.
- Erik F Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179. Association for Computational Linguistics, 1999.
- Hans Jürgen Sasse. Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state? *Linguistic Typology*, 6(2):199–271, 2002. ISSN 14300532. doi: 10.1515/lity.2002.007.
- Roser Saurí. Annotating temporal relations in Catalan and Spanish TimeML annotation guidelines. Technical Report BM 2010-04, Barcelona Media, 2010.
- Roser Saurí and James Pustejovsky. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268, 2009. ISSN 1574020X. doi: 10.1007/s10579-009-9089-9.

BIBLIOGRAPHY

- Guido Sautter, Klemens Böhm, Frank Padberg, and Walter Tichy. Empirical evaluation of semi-automated XML annotation of text documents with the GoldenGATE editor. *Research and Advanced Technology for Digital Libraries*, pages 357–367, 2007.
- Frank Schilder and Christopher Habel. From Temporal Expressions to Temporal Information : Semantic Tagging of News Messages. In *ACL01 workshop on temporal and spatial information processing*, volume at-time, pages 65–72, Toulouse, France, 2001. Association for Computational Linguistics. doi: 10.3115/1118238.1118247. URL <http://www.informatik.uni-hamburg.de/WSV/hp/schilder/ACL01WS.pdf>.
- Frank Schilder and Christopher Habel. Temporal information extraction for temporal question answering. In *New Directions in Question Answering*, pages 35–44. AAAI, 2003.
- Mary Suzanne Schriber. Women’s Place in Travel Texts. *Prospects*, 20: 161–179, 1995.
- Roxane Segers, Marieke Van Erp, Lourens Van Der Meij, Lora Aroyo, Guus Schreiber, Bob Wielinga, Jacco van Ossenbruggen, Johan Oomen, and Geertje Jacobs. Hacking history: Automatic historical event extraction for enriching cultural heritage multimedia collections. In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP’11)*, pages 26–29, 2011.
- Andrea Setzer. *Temporal information in newswire articles: an annotation scheme and corpus study*. PhD thesis, University of Sheffield, 2001.
- Ryan Shaw. A semantic tool for historical events. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 38–46, 2013.

BIBLIOGRAPHY

- Ryan Shaw, Raphaël Troncy, and Lynda Hardman. LODE: Linking Open Descriptions of Events. *ASWC*, 9:153–167, 2009.
- Ryan Benjamin Shaw. *Events and periods as concepts for organizing historical knowledge*. University of California, Berkeley, 2010.
- F Simiand. *Annales. Histoire, Sciences Sociales, France: EHESS*, 15(1): 83–119, 1960.
- Guillermo Soberon, Lora Aroyo, Chris Welty, Oana Inel, Manfred Overmeen, and Hui Lin. Content and Behaviour Based Metrics for Crowd Truth. In *CrowdSem*, pages 45–58, 2013.
- Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. From light to rich ERE: annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 89–98, Denver, Colorado, USA, 2015. Association for Computational Linguistics.
- Zhiyi Song, Ann Bies, Stephanie Strassel, Joe Ellis, Teruko Mitamura, Hoa Dang, Yukari Yamakawa, and Sue Holm. Event Nugget and Event Coreference Annotation. In NaacL Hlt, editor, *4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 37–45, 2016.
- Manuela Speranza and Rachele Sprugnoli. Annotation of Temporal Information on Historical Texts: a Small Corpus for a Big Challenge. In *Proceedings of the "Formal Representation and Digital Humanities" workshop*, IN PRESS.
- Richard C Sprinthall. *Basic statistical analysis*. Allyn & Bacon, 2003.
- Rachele Sprugnoli. L’annotazione dei nomi di evento per il trattamento automatico della lingua. Master’s thesis, University of Pisa, 2012.

BIBLIOGRAPHY

- Rachele Sprugnoli. Two days we have passed with the ancients...: a Digital Resource of Historical Travel Writings on Italy. In *Proceedings of the 7th AIUCD Conference*, 2018.
- Rachele Sprugnoli and Alessandro Lenci. Crowdsourcing for the identification of event nominals: an experiment. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014. ISBN 978-2-9517408-8-4.
- Rachele Sprugnoli and Sara Tonelli. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4):485–506, 2017.
- Rachele Sprugnoli, Giovanni Moretti, Sara Tonelli, and Stefano Menini. Fifty years of European history through the Lens of Computational Linguistics: the De Gasperi Project. *ITALIAN JOURNAL OF COMPUTATIONAL LINGUISTICS*, 2(2):89–100, 2016.
- Rachele Sprugnoli, Tommaso Caselli, Sara Tonelli, and Giovanni Moretti. The content types dataset: A new resource to explore semantic and functional characteristics of texts. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, volume 2, 2017a. ISBN 9781510838604.
- Rachele Sprugnoli, Giovanni Moretti, Luisa Bentivogli, and Diego Giuliani. Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation*, 51(2):283–317, 2017b.
- Rachele Sprugnoli, Sara Tonelli, Giovanni Moretti, and Stefano Menini. A little bit of bella pianura: Detecting code-mixing in historical English travel writing. In *CEUR Workshop Proceedings*, volume 2006, 2017c.

BIBLIOGRAPHY

- Rachele Sprugnoli, Giovanni Moretti, and Sara Tonelli. LOD Navigator: Tracing Movements of Italian Shoah Victims. In *Proceedings of EHRI workshop Data Sharing, Holocaust Documentation, Digital Humanities*, IN PRESS.
- Manfred Stede and Florian Kuhn. Identifying the content zones of German court decisions. In *International Conference on Business Information Systems*, pages 310–315. Springer, 2009.
- Dan Stone. *Holocaust Spaces*. University of Chicago Press, 2016.
- Stephanie Strassel. Topic detection and tracking annotation guidelines: Task definition to support the tdt2002 and tdt2003 evaluations in english, chinese and arabic. Technical report, Linguistic Data Consortium, 2005.
- Jannik Strötgen, Ayser Armiti, Tran Van Canh, Julian Zell, and Michael Gertz. Time for More Languages: Temporal Tagging of Arabic, Italian, Spanish, and Vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):1–21, 2014.
- Rudi Studer, V Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2): 161–197, 1998.
- William F. IV Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2(1):143–154, 2014. ISSN 2307-387X. URL <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/305>.
- Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American*

BIBLIOGRAPHY

- Medical Informatics Association*, 20(5):806–813, 2013. ISSN 10675027. doi: 10.1136/amiajnl-2013-001628.
- Rosemary Sweet. *Cities and the Grand Tour: the British in Italy, c. 1690-1820*. Number 19. Cambridge University Press, 2012.
- Leonard Talmy. Fictive motion in language and ception. *Language and space*, 21:1–276, 1996.
- Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1493–1502. Association for Computational Linguistics, 2009.
- The European Union. Definition of the Europeana Data Model elements. Technical report, Europeana, 2012.
- Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Tech. Rep*, 2012.
- Katrin Tomanek and Fredrik Olsson. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48. Association for Computational Linguistics, 2009.
- Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéal. LIMSI-COT at SemEval-2017 Task 12: Neural Architecture for Temporal Information Extraction from Clinical Narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 597–602, Vancouver, Canada, 2017. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S17-2098>.

BIBLIOGRAPHY

- Jouni Tuominen. Bio CRM: A Data Model for Representing Biographical Information for Prosopography. Version 2016-08-19. Technical report, Bringing Lives to Light: Biography in Context Project, 2016. URL <https://seco.cs.aalto.fi/projects/biographies/biocrm-2016-08-19.pdf>.
- Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 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, pages 1–9, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.
- Chiel Van Den Akker, Lora Aroyo, Agata Cybulska, Marieke Van Erp, Peter Gorgels, Laura Hollink, Cathy Jager, Susan Legene, Lourens van der Meij, Johan Oomen, et al. Historical event-based access to museum collections. In *Proceedings of the First International Workshop on Recognising and Tracking Events on the Web and in Real Life (EVENTS2010)*, Athens, Greece, 2010.
- Johan Oomen Roxane Segers Chiel van den Akker Lora Aroyo Geertje Jacobs Susan Legêne Lourens van der Meij Jacco van Ossenbruggen Van Erp Marieke and Guus Schreiber. Automatic Heritage Metadata Enrichment with Historic Events. In *Museums and and the Web*, 2011.
- Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136, 2011.
- Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. Hope and fear: Interpreting perspectives by integrating sentiment and

BIBLIOGRAPHY

event factuality. In Iceland Reykjavik, editor, *Proceedings of LREC 2014*, pages 26–31. European Language Resources Association (ELRA), 2014.

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy Chapman. BluLab: Temporal Information Extraction for the 2015 Clinical TempEval Challenge. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 815–819, Denver, Colorado, 2015. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/S15-2137>.

Zeno Vendler. Verbs and Times. In *The Philosophical Review*, volume 66, page 143. Cornell University Press, Ithaca, NY, 1957. ISBN 0031-8108. doi: 10.2307/2182371. URL <http://www.jstor.org/stable/2182371?origin=crossref>.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. SemEval-2007 Task 15 : TempEval Temporal Relation Identification. In *Computational Linguistics*, number June, pages 75–80, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W/W07/W07-2014>.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. SemEval-2010 task 13: TempEval-2. In *ACL 2010 - SemEval 2010 - 5th International Workshop on Semantic Evaluation, Proceedings*, number July, pages 57–62, Uppsala, Sweden, 2010. Association for Computational Linguistics. ISBN 1932432701.

Piek Vossen, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shukai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, and Others. KYOTO: a System for Min-

BIBLIOGRAPHY

- ing, Structuring and Distributing Knowledge across Languages and Cultures. In *Proceedings of LREC*, 2008a.
- Piek Vossen, Isa Maks, Roxane Segers, and Hennie VanderVliet. Integrating Lexical Units, Synsets and Ontology in the Cornetto Database. In *Proceedings of LREC*, 2008b.
- Piek Vossen, German Rigau, Luciano Serafini, Pim Stouten, Francis Irving, and Willem Van Hage. NewsReader: recording history from daily news streams. In Iceland Reykjavik, editor, *Proceedings of LREC 2014*. European Language Resources Association (ELRA), 2014.
- Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines*, pages 40–49, 2015.
- Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, and Others. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85, 2016.
- Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31, 2013.
- Egon Werlich. *A text grammar of English*. Quelle & Meyer, 1976.
- Bao Quoc Ho Xuan Quang Pham, Minh Quang Le. A Hybrid Approach for Biomedical Event Extraction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 121–124, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

BIBLIOGRAPHY

- Yadollah Yaghoobzadeh, Gholamreza Ghassem-Sani, Seyed Abolghasem Mirroshandel, and Mahbaneh Eshaghzadeh. ISO-TimeML event extraction in persian text. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, volume 1, pages 2931–2944, Mumbai, India, 2012. Association for Computational Linguistics.
- Vanni Zavarella and Hristo Tanev. FSS-TimEx for TempEval-3: Extracting Temporal Information from Text. In *Proceedings of SemEval 2013*, pages 58–63, Georgia, USA. Association for Computational Linguistics., 2013. Atlanta.
- Matthew D Zeiler. Adadelata: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
- Amir Zeldes. The gum corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017.
- Deyu Zhou, Xuan Zhang, and Yulan He. Event extraction from Twitter using Non-Parametric Bayesian Mixture Model with Word Embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 808–817, 2017.

Appendices

Appendix A

Case Study: the Shoah Ontology and the Events of Movement

In this Appendix we report a work carried out in collaboration with Giovanni Moretti and Sara Tonelli and presented during the European Holocaust Research Infrastructure (EHRI) workshop on “Data Sharing, Holocaust Documentation, Digital Humanities”, held in Venice in June 2017. The application described below (i.e., *LOD Navigator*) is an example of the exploitation of an ontology created in the Digital Humanities, directly connected to the ones illustrated in Section 3.1.2. The description of the *LOD Navigator* is published in [Sprugnoli et al., IN PRESS].

A.1 Tracing Movements of Italian Shoah Victims

A database with information on Italian Shoah victims has been developed at the Contemporary Jewish Documentation Center in Milan (CDEC) and was made freely available as Linked Open Data (LOD). The database was built starting from information collected in Fargion [Fargion, 1991] and is accessible through a web portal (the *CDEC Digital Library*¹) and a

¹<http://digital-library.cdec.it/cdec-web/>

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

SPARQL endpoint². In order to provide a novel way to navigate this database taking advantage of the available high-quality records, we developed the *LOD Navigator*, a system that allows users to explore the trajectories of victims during their persecution both at micro and at macro level. The goal of the application is to enhance the value of the original database with a user-friendly interface focusing on an important research field in Shoah studies, i.e. the geographies of the Holocaust.

Information necessary for implementing the *LOD Navigator* was collected using the SPARQL endpoint: this information includes biographical data together with details about the persecution and deportation of each victim. The places of birth, arrest, detention, deportation to a Nazi camp, transfer, and return after liberation (if available) were then semi-automatically georeferenced and associated with the corresponding date. We then considered a movement as a trajectory from one georeferenced place (associated with a dated event) to another georeferenced place. All the movements are then displayed in the *LOD Navigator* through an interactive interface made freely available as a standalone tool³.

We decided to focus our attention on movements because the Holocaust was characterized by many spatial processes. Concentration, deportation, dispersal, dislocation are all geographical components involved in the implementation of the Nazis' genocidal policy. Plotting data about these spatial processes on a map and giving them a temporal dimension allows the identification of spatio-temporal patterns at the macro-level but also at the micro-level, to reconstruct individual experiences.

²<http://dati.cdec.it/>

³<http://dh.fbk.eu/technologies/lod-navigator>

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

A.1.1 Related Work

The ability to move groups of people separating them from their original social context is a crucial strategy in every genocide [Dadrian, 2004]. In the case of Nazism, the genocidal policies were characterised by radical spatial acts euphemistically represented in Schutzstaffel’s bureaucratic jargon with terms such as *Auswanderung* (emigration), *Aussiedlung* (re-settlement) and *Wohnsitzverlegung* (change of residence) [Hilberg, 1985]. The literature reports a number of works dealing with this territorial dimension of the Holocaust and thus belonging to what Knowles et al. call the “spatial turn in Holocaust studies” [Knowles et al., 2015].

Theoretical and historiographical aspects of Nazism seen as a spatial project are discussed in the papers collected in Giaccaria and Minca [Giaccaria and Minca, 2016]. In particular, Stone [Stone, 2016] highlights how Holocaust affected the entire continent: it was not just a matter related to specific places such as extermination camps, but the violence was widespread in both small and big sites in every nation.

In [Beorn et al., 2009] and [Knowles et al., 2014] the focus is on the use of GIS (Geographic Information System) and geovisualisation as means to rethink the Holocaust at different levels of analysis. The former, for example, presents a prototype visualisation of the journeys of eight Hungarian Jews, while the work by Giordano and Holian [Giordano and Holian, 2014] analyses the patterns of arrests of Jews in Italy from a spatio-temporal perspective, using “Libro della Memoria” as their source of data. This last analysis is paired with an interactive visualisation showing the distribution of arrests per month during the period 1943-1945⁴. This representation does not take into consideration all the persecution stages happened after the arrest. In addition, not all victims are included in the analysis: for

⁴http://web.stanford.edu/group/spatialhistory/cgi-bin/site/viz.php?id=383&project_id=0

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

example, those killed in massacres or arrested outside Italy are excluded.

With respect to these examples of previous work, our application differs in at least 3 aspects: (i) it follows the lives of the largest possible number of victims, so as to provide information representative of the phenomenon; (ii) it represents all the major events related to the persecution of Jews; (iii) it offers novel insight not only by aggregating data according to several categories, but also by allowing the close reading of individual stories.

A.1.2 Workflow

To develop the *LOD Navigator*, we performed 4 main tasks: (i) we manually analysed the available data, (ii) we extracted the data that were interesting for our aim, (iii) we manipulated them to remove inconsistencies and add missing information. Finally, (iv) we implemented the data visualisation functionalities of the application. These phases are detailed in the next subsections.

Data Observation First of all, we manually analysed the data available in the CDEC digital library and the corresponding RDF browser to select the information we should focus on to build the application. In the RDF browser, biographical information is given for each person, identified by a unique ID, together with information about his/her persecution. This information is structured on the basis of the Shoah domain ontology that formally describes concepts and relationships characterizing the process of persecution of Jews in Italy between 1943 and 1945 [Brazzo and Mazzini, 2015, 2017]. In particular, the ontology Class called *Persecution* describes the arrest, detention, deportation to a Nazi camp, the transfer to another camp but also the liberation and massacre. This class is related to the *Person* class that includes properties connected to biographical information, such as the date and place of birth and death. Other information, for

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

example family relations between victims, was present in the database but was not considered relevant in this phase of the development.

Data Extraction We queried the SPARQL endpoint to retrieve all the information we needed and we chose to obtain results in CSV format. In this way we extracted the content of 26 properties (see Figure A.1) for 9,042 people identified as Italian victims of the Shoah. These properties include basic personal information such as gender and place of birth and death, but also details on the different events related to Shoah in each person’s biography, for example the date and place of arrest, the date of transfer to the Nazi camp and the type of death. Overall, we selected all properties necessary to trace the personal trajectories of the victims along a timeline.

1. ID	14. shoahSurvivor
2. name	15. arrestDate
3. gender	16. arrestPlace
4. date_of_birth	17. departureDate
5. place_of_birth	18. convoyDeparturePlace
6. date_of_death	19. convoyNumber
7. place_of_death	20. convoyArrivalDate
8. massacreDate	21. arrivalDate
9. massacrePlace	22. toNaziCamp
10. deathDescriptionIntegration	23. naziCampTransfertDate
11. position	24. toNaziCampTranfert
12. abstract	25. liberationReturnDate
13. deathDescription	26. liberationReturnPlace

Figure A.1: Properties whose content was extracted from CDEC LOD dataset.

Data Manipulation and Integration In the third step, we performed a semi-automatic check of the data formats to fix inconsistencies. Since the information had been manually recorded, possibly by different persons, in some cases data formats and conventions were not homogeneous. Besides, we decided to simplify some information associated with each biography and

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

add details that can help the user’s navigation.

A particularly challenging case was the format of dates, which we found in 14 different variants including at least the year (e.g., *19250315*, *1900*, *16/06/1944*, *1944.02.26*). These date versions have been converted, using a Python script, in a unique format: YYYY-MM-DD. We also found 8 different conventions to express the lack of temporal information, such as *0* or *?*. In this case, we could not give a temporal anchor to the corresponding event, therefore we had to remove the event from the database.

We also decided to modify people occupations by clustering the available options into coarse-grained categories to simplify navigation for end users. While in the LOD database we found 168 occupations, we observed that some of them were only little different, for example “calzettaia” / *female hosier* and “calzettaio” / *male hosier*. So we used Wikipedia classification of occupation types⁵ and we selected 27 coarse-grained categories, onto which we mapped the original ones. In this way “calzettaia / calzettaio / calzolaio” were clustered all under the category “Craft occupations”. For all victims without an associated occupation, the *unknown* category was added.

In case the `place_of_death` property had no value, we automatically extracted the places mentioned in the free-text descriptions of the field `deathDescriptionIntegration`.

This was performed using *The Wiki Machine* [Palmero Aprosio and Giuliano, 2016], a system that links the concepts mentioned in a document to the corresponding Wikipedia page describing them. By selecting only the pages referring to a place, we automatically identified geographical mentions. For instance, giving as input the sentence “Ucciso in tentativo di fuga a Milano” / *Killed in an attempt to escape in Milan*, the tool links “Milano” to the corresponding Wikipedia page, and annotates the word as a

⁵https://en.wikipedia.org/wiki/Category:Occupations_by_type

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

location, more specifically as an administrative region, as shown in Figure A.2.

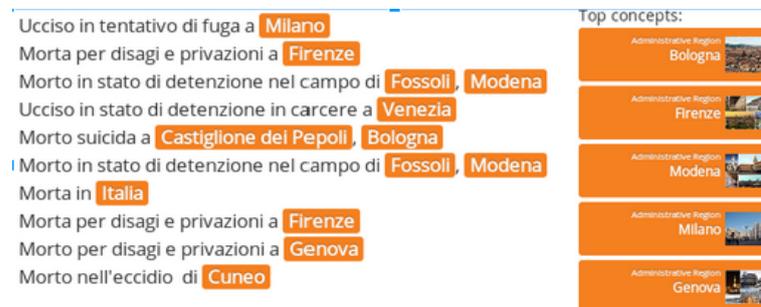


Figure A.2: Example of output of the Wiki Machine.

Another integration we performed semi-automatically was adding the country of origin of each victim. This information was obtained by looking up the `abstract` property, which always contains a sentence such as: “She was born in France...”.

Since the final goal of *LOD Navigator* is to display the places on a map, another crucial step involved georeferencing place names, that is finding the coordinates of locations. In total we found 1,493 unique places in the data and we used Nominatim⁶ to automatically retrieve their latitude and longitude. In 15% of the cases this automatic process failed and we had to manually correct wrong coordinates (for example, Nominatim locates Siena in China) or we had to georeference places by hand because Nominatim could not find any coordinate. Sometimes this was due to the presence of some non-standard spelling but through spelling normalization it was possible to map different versions to the correct name and to precise coordinates (e.g. spelling variant: Gross-Meseritz – standard spelling: GroßMeseritsch). For other locations, we could only make an approximation by using latitude and longitude of the country they should belong to, inferring this information from the content of other fields. For example, we could not locate Choumbla, but it was possible to associate it to

⁶<https://nominatim.openstreetmap.org/>

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

the coordinates of Bulgaria, taking this information from the content of the **abstract** property: *Menachem Levi, son of Haim Levi and Veneziana Benveniste. He was born in Bulgaria, in Choumbla, on July 23 1876.*

At the end of the step, all retrieved data were corrected and harmonised, and they were converted in JSON format. Each victim was associated with biographical information and the list of his/her movements.

Visualization The last step involved building a tool to support data visualization and navigation. To this end we adapted and expanded the interactive interface of **RAMBLE ON** [Menini et al., 2017].

A.1.3 Quantitative and Qualitative Data Analysis

The *LOD Navigator* contains information about 8,712 victims (4,470 males, 4,239 females, and 3 of unknown gender) that, according to the analysis described in Section A.1.2, were found to have at least one dated movement between two georeferenced places. This means that the application displays the lives of 96.3% of all the victims recorded in the CDEC LOD dataset. Most of them (89.8%) do not have a specified occupation: among the others, the most common category is *Sales occupations* with 407 people having jobs such as salesperson, peddler and shop assistant. As for the country of origin, victims were born in 37 different nations: Italy is the most represented country (4,276 people) however the remaining 36 countries cover 49% of all the victims. This shows that, although we deal with Italian Shoah, around half of the victims were either arrested, detained or dead in Italy, but were originally from other countries. In many cases, for example for many French and Eastern European Jews, Italy was the country where they fled in the hope to escape from Nazi threat while others lived in Italian possessions that, after the Armistice of Cassibile, were occupied by German troops.

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

Victims' fate is described by the value of `death_description` property, used as a filter in the application. We report in Table A.1 some statistics extracted with the help of *LOD Navigator*: sadly, the great majority of victims in the database (84.2%) died in an extermination camp, and only 12% survived the Shoah.

<code>death_description</code>	QUANTITY
Dead in extermination camp	7,333
Survivor of Shoah	1,037
Dead in massacre	202
Dead in custody	70
Unknown	28
Dead en route to camp	15
Committed suicide	10
Dead of hardships and privations	6
Killed in escape attempt	6
Missing	6
Killed during arrest	1
TOTAL	8,712

Table A.1: Division of victims on the basis of the `death_description` property.

We also show in Table A.2 the locations displayed in the application interface that are more frequently associated with specific events. Each row in the table can be explained by connecting it to the history of Jewish communities in Italy and in Europe in the XX Century. For example, many victims were born in Rome, Trieste and Venezia, because at the time they had the largest Jewish communities in Italy. Rhodes was also the birthplace of many victims, because it had been under Italian control since 1912 and in the 1920s the local Jewish community was very important, including around one-third of the total population. Rhodes is also the place witnessing the highest number of arrests (1,758 out of 7,960), confirming that the Shoah led to the deportation and the death of most of

A.1. TRACING MOVEMENTS OF ITALIAN SHOAH VICTIMS

the community members.

EVENTS	#	TOP 5 PLACES
Birth	8,340	Roma (1,526) - Rhodes (1,524) - Trieste (466) - Vienna (205) - Venezia (200)
Arrest	7,960	Rhodes (1,758) - Roma (1,712) - Trieste (509) - Borgo San Dalmazzo (329) - Milano (241)
Detention	7,873	Fossoli (2,699) - Rhodes (1,712) - Roma (1,029) - Milano (872) - Trieste (661)
To Nazi Camp	7,893	Auschwitz (7,338) - Bergen Belsen (405) - Ravensbrueck (89) - Flossenburg (27) - Buchenwald (25)
Nazi Camp Transfer	821	Vittel (107) - Monowitz (86) - Biberach (75) - Buchenwald (59) - Mauthausen (56)
Death	4,332	Auschwitz (3,743) - Roma (81) - Monowitz (57) - Lago Maggiore (50) - Flossenburg (38)
Return after Liberation	136	Roma (86) - Milano (32) - Torino (7) - Tripoli (4) - Livorno (2)
TOTAL	37,355	

Table A.2: Number of events dated and georeferenced in the *LOD Navigator* together with the five most frequent locations for each event.

Despite being a small city, also Borgo San Dalmazzo is among the places where most of the arrests were carried out. This is because it hosted a Nazi concentration camp, where foreign-born Jews from France were arrested trying to escape the Vichy regime. Also Fossoli is just a small village in Emilia Romagna, but it was the place of detention for most of the recorded victims because there was a transit camp where Jews were detained before being sent to Auschwitz. The latter is the camp where most of the victims in CDEC database were deported and died. In addition to other two extermination camps (i.e., Monowitz and Flossenburg), many Jews died in massacres as in Rome (Fosse Ardeatine) and on the Lake Maggiore. Information about the return after liberation is available only

A.2. APPENDIX SUMMARY

for 136 people. The proportion between Shoah victims that were born in Rome and those that returned to the city after liberation (1,526 vs. 86) shows, as an eloquent example, the impact of this tragedy on local Jewish communities.

A.2 Appendix Summary

In this Appendix we present the *LOD Navigator* whose contribution is manifold: it provides for the first time an interactive system, through which part of the data collected by CDEC can be browsed, searched and visually displayed on a map. This can benefit the community of researchers interested in studying the Shoah but also the Jewish community and the broad public. Thus this work constitutes an example of digital history application. We believe that the main innovative idea of the *LOD Navigator* is to visually track movements that, when available only in LOD or plain text, do not give the possibility to get an overview or to interact with the data. It addresses the needs of researchers and scholars but also of the families involved in deportations and all the Jewish community to support their effort in reconstructing stories of families and losses during the Shoah. Besides, the navigation system is data-independent and can be used to view other trajectories by uploading a simple JSON file containing georeferenced places and dates, thus providing a useful tool to (digital) history scholars at large: for example, in Menini et al. [2017], a preliminary version of the interface was adopted to trace the motion trajectories automatically extracted from Wikipedia biographies with the aim of providing important data for the analysis of culture and society.

A.2. APPENDIX SUMMARY

Appendix B

Questionnaire: What is an Event in History?

This Appendix contains the content of the questionnaire described and discussed in Chapter 4. The questionnaire was created using Google Form and consists of 18 questions: an asterisk marks required questions. We circulated it both in English (see Section B.1) and in Italian (see Section B.2).

B.1 English Questionnaire

B.1.1 Introduction

We are conducting a questionnaire about the concept of “event” for historians and the interest towards the use of Natural Language Processing tools to support historical investigations. Your answers will help us find out how historians define events encoded in late modern and contemporary historical texts and will contribute to the development of a content analysis tool tailored to historical research needs. This questionnaire is part of an ongoing PhD project running at the Digital Humanities group at Fondazione Bruno Kessler, Trento, Italy: <http://dh.fbk.eu/>. For questions about this questionnaire, please write an email to dh-survey@fbk.eu. The full

B.1. ENGLISH QUESTIONNAIRE

questionnaire consists of 18 questions. It should take you no more than 20 minutes to respond to all questions. All the data we receive will be anonymous, unless you decide to tell us your name and e-mail address so that we can send you news of future surveys, experimentation and findings. We will not use the addresses we gather for any other purpose, although we may use comments submitted in this questionnaire but in an anonymous way. Thank you for taking the time to respond!

B.1.2 Part 1 - Events in historical texts

1. "Today, once again, the independence of the Western Hemisphere is menaced from abroad." *

Could you please list all the words / expressions conveying events in the previous sentence? Please separate each word / expression with a comma (,). If for you there are NO events in the sentence, please write *NO*.

2. "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." *

Could you please list all the words / expressions conveying events in the previous sentence? Please separate each word / expression with a comma (,). If for you there are NO events in the sentence, please write *NO*.

3. "I think we can work that out with the advice of the Ways and Means Committee." *

Could you please list all the words/expressions conveying events in the previous sentence? Please separate each word/expression with a comma (,). If for you there are NO events in the sentence, please write *NO*.

B.1. ENGLISH QUESTIONNAIRE

4. Imagine you are analysing a textual source: what are the most important properties you look at in order to understand if a word (or a set of word) expresses a relevant event? *

Rate the level of importance of each of the following properties you take into consideration when dealing with expressions encoding events within historical texts.

	Not important	Somewhat important	Very important	Don't know
Participants (e.g. important vs common people)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Type (e.g. conflict events "massacre", reporting events "declare", cognitive events "believe")	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Duration / period of occurrence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Frequency of occurrence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Agency (e.g. caused by humans; not caused by humans)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Impact (e.g. on humans, animals or nature)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cause of the event	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Preceding and consequent event(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Predictability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Public perception	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Factuality (e.g. certain events versus probable events)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5. On the basis of your historical research practice, how would you define an event? *

B.1. ENGLISH QUESTIONNAIRE

6. Which of the following annotations fits better with your notion of events in historical texts? *

In the following two images you see the same sentence where events are annotated and highlighted in yellow following two different strategies. In the first one only single words can be annotated as events (see *show* versus *showed up*) and states/conditions such as *independence* are not taken into consideration.

- Annotation #1
- Annotation #2
- Neither of the two

After the key African state of Guinea, now voting with the Soviet Union in Communist foreign policy, after it

1 gained its independence, a Russian Ambassador showed up the next day. Our Ambassador did not show up for 9 months.

After the key African state of Guinea, now voting with the Soviet Union in Communist foreign policy, after it

2 gained its independence, a Russian Ambassador showed up the next day. Our Ambassador did not show up for 9 months.

B.1.3 Part 2 - Natural Language Processing

1. Do you know what Natural Language Processing is? *

- No, but I'd like to know more
- No, I'm not interested
- I've seen this expression, but don't really understand it
- Yes, I consider myself an expert
- Yes, I use (or I've used) Natural Language Processing methods in my research

B.1. ENGLISH QUESTIONNAIRE

- Yes, but I have never used Natural Language Processing methods in my research
2. Which of the below Natural Language Processing techniques do you know or do you use in your research? (click all that apply) *
- Tokenization
 - Sentence breaking
 - Part-of-speech tagging
 - Lemmatisation
 - Syntactic analysis
 - Named Entity Recognition (NER)
 - Relationship extraction
 - Sentiment analysis
 - Text simplification
 - Topic modeling
 - Other: (please specify)
3. If a tool that helps historians to discover and analyse events, temporal expressions and temporal relations within texts were made available to you, would you be interested in it? *
- Yes, but I'd need training
 - Yes
 - Perhaps
 - No, I'm not interested / I dont need such tool
4. Could you please briefly explain the motivation behind your previous answer? *

B.1. ENGLISH QUESTIONNAIRE

1. In which country do you live? *

2. What is your age? *

- 20-30 years old
- 30-40 years old
- 40-50 years old
- 50-60 years old
- > 60 years old

3. What is your gender? *

- Female
- Male
- I prefer not to answer this question

4. What is your current academic position? *

- PhD candidate
- Early career researcher (between 1 and 10 years post-doctoral)
- Senior researcher
- Professor
- Other: (please specify)

5. How would you define your field of research? (click all that apply) *

- History of the arts
- Biography and psychohistory
- Diplomatic history
- Economic history

B.2. ITALIAN QUESTIONNAIRE

- Intellectual history
- Military history
- Political history
- History of science
- Social and cultural history
- Gender history
- Other: (please specify)

6. In which languages are you interested in your research? (click all that apply) *

- Italian
- English
- French
- German
- Spanish
- Greek
- Latin
- Arabic
- Other: (please specify)

B.2 Italian Questionnaire

The Italian questionnaire was the translation of the English one documented in the previous Section. Here we report only the questions belonging to the first part and involving examples taken from historical documents.

B.2. ITALIAN QUESTIONNAIRE

B.2.1 Parte 1 - Eventi nei testi storici

1. “Man mano che avanzavano, i soldati andavano a prender posto nei vagoni del lunghissimo convoglio.” *

Puoi per favore elencare tutte le parole/espressioni che esprimono eventi nella frase precedente? Separa ogni parola/espressione con una virgola (,). Se secondo te questa frase non contiene eventi, scrivi *NO*.

2. “Invece la guerra non è ancora finita e la pace sembra ancora lontana.” *

Puoi per favore elencare tutte le parole/espressioni che esprimono eventi nella frase precedente? Separa ogni parola/espressione con una virgola (,). Se secondo te questa frase non contiene eventi, scrivi *NO*.

3. “Ella voglia la prego aggiungere che ci rendiamo perfettamente conto della delicatezza della questione che incide sui rapporti stessi fra i grandi alleati.” *

Puoi per favore elencare tutte le parole/espressioni che esprimono eventi nella frase precedente? Separa ogni parola/espressione con una virgola (,). Se secondo te questa frase non contiene eventi, scrivi *NO*.

4. Quale delle seguenti annotazioni si avvicina alla tua nozione di evento nei testi storici? *

Nelle due immagini che seguono puoi vedere la stessa frase in cui gli eventi sono annotati ed evidenziati in giallo seguendo due diverse strategie. Nella prima sono annotate solo parole singole (vedi *fare* versus *fare sul serio*) e gli stati/condizioni come *ottimismo* non sono presi in considerazione.

- Annotazione #1
- Annotazione #2
- Nessuna delle due

B.2. ITALIAN QUESTIONNAIRE

1 A Londra si caricano forse le tinte per convincere la Germania che la Russia vuol proprio fare sul serio ; a Berlino incomincia a scemare l'ottimismo degli altri giorni , da Vienna non si è comunicato alla stampa nessuna notizia

2 A Londra si caricano forse le tinte per convincere la Germania che la Russia vuol proprio fare sul serio ; a Berlino incomincia a scemare l'ottimismo degli altri giorni , da Vienna non si è comunicato alla stampa nessuna notizia ,