EEA Financial Mechanism 2009-2014

Deliverable D4.2

Report on the Event Database

Project Number: EOX GR07/3712

Project Title: Examining the phenomenon of xenophobia in Greece during the economic crisis: A computational perspective

Subproject 2: Computational methods and techniques for analyzing xenophobia in Greece (ΣΑΕ”013/8”)

ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ,
ΕΡΕΥΝΑΣ ΚΑΙ
ΘΡΗΣΚΕΥΜΑΤΩΝ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ
ΤΕΧΝΟΛΟΓΙΑΣ

ICELAND
LIECHTENSTEIN
NORWAY
eea grants

ATHENA
Research & Innovation
Information Technologies

Deliverable 4.2
Report on the Event Database

Date:	December 2016
Authors:	Konstantina Papanikolaou, Dr. Haris Papageorgiou, Maria Pontiki, Dimitris Pappas, Dimitris Gkoumas, Penny Labropoulou, Maria Gavriilidou, Stelios Piperidis
Dissemination level:	Public
Version:	1.0
Keywords:	Event Analysis, Protest Events, Physical Violence, News Agencies, Xenophobia
Description:	This Deliverable reports on the Event Database which was generated in the context of the examination of Xenophobia in Greece during the last 20 years. More specifically, two major Event Categories were analysed, namely Physical Attacks and Protests, involving specific ethnic groups which were defined. The Event Database aspires to contribute to answering the research questions posed by the current project.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The aim of the present project was to examine Xenophobia in Greece during the last two decades. Moreover, it intended to investigate whether the economic crisis that has ensued in the country since 2008, has affected xenophobic sentiments within the Greek society, and consequently to confirm or reject the perception that xenophobia generally escalates under circumstances of severe economic distress.

According to the literature, Xenophobia as a phenomenon is characterized mainly by prejudices against what is considered as “other” as well as by emotions like fear and hatred. The scope of this research was to explore possible realizations of such emotions and prejudices, namely the way they are reflected to real world events. To this end, a large Event Database was designed and implemented, capturing physical violence along with protests. It is our claim that this is one of the ways prejudices and xenophobic sentiments can be expressed.

The present Deliverable 4.2. is to describe the above-mentioned Event Database along with the methodology which was implemented for its generation. The report is structured as follows: first, an overview of the literature is presented, concerning Event Extraction as a task as well as its relation to Political and Social Sciences (Section 2). After that, every step of the Event Extraction methodology which was followed is described in details (Section 3). Next, the Results of the analysis are presented, namely the Event Database itself is delineated along with some illustrative visualizations (Sections 4 and 5). Finally, the Deliverable concludes with a brief summarization of the process, the results and the issues that supervened throughout the analysis, together with some thoughts about future work (section 6).

2. Literature

In the context of Information Extraction (IE), several frameworks have been proposed for Event Extraction. Two of the most influential are the Probabilistic Event Model and the Structural Event Model. The first one was put forward within the framework of the Topic Detection and Tracking (TDT) study (Allan et al. 1998) and especially the Retrospective news Event Detection (RED) task (Li et al. 2005). This study defined an event as "some unique thing that happens at some point in time". Four types of information are used to represent events within news articles: *who* (person), *when* (time), *where* (location) and *what* (keywords) (Wang & Zhao 2012). Event extraction is regarded mostly as a classification, clustering problem. Given a stream of news data, the task is to segment into stories and classify each story as referring to an event or not. This is considered an unsupervised learning task, as it uses no labeled training examples.

The Structural Event Model was proposed in the context of Message Understanding Conferences (MUC). The scope of these challenges was to complete all the slots within specific event templates. Thus, Event is considered a structure of certain information types. Consequently, specific event (scenario) templates (Chinchor & Marsh, 1998) are defined and, given an input text, all the relevant information needed to be identified and then relate the entities to the event they are involved to. The approach adopted by this work is closer to the structural model, in the sense that specific event types are coded and event templates comprising certain information types are determined. Therefore, the task is to first identify each information type within a given text and then link them to fill the slots within an event template.

There are various views regarding what exactly constitutes an event and how, through which linguistic structures it is expressed in data (Pustejovsky 2000). Moreover, many research efforts have focused on detecting and identifying events (Allan et al. 1998, Filatova & Hatzivassiloglou 2003, Filatova & Hovy 2001, Yang et al. 1999). All the above-mentioned approaches aim at developing applications which can automatically extract events and recognize the spatial and temporal relations connecting them.

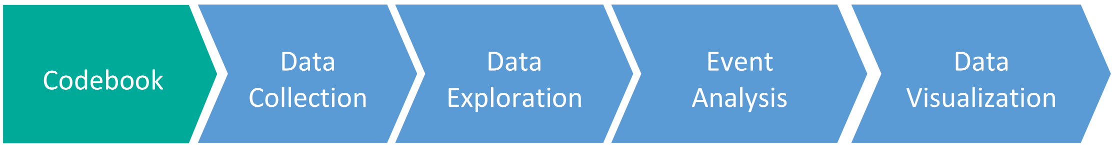
Event extraction for political and social science has been a long-standing topic, dating back to hand coding data. Work on automatic annotation started within the KEDS/TABARI project (Shrodt, Shannon and Weddle 1994). Evaluations have shown that hand coded and automatic events coding show comparable performance (King and Lowe 2003). Several coding schemes have been developed since, including the

IDEA (Bond et al. 2003) and ICEWS (O' Brien 2012). One of the most renown and influential frameworks for event extraction is CAMEO (Gerner et al. 2002), which is still used by the ongoing GDELT project (Leetaru and Shrodt 2013). All these efforts have focused on news data, that have traditionally been the main events' source. Our codebook, follows the same principles with a linguistically-driven implementation. Protest Events Analysis has been a central issue in the context of Political and Social sciences (Wueest, Rothenhäusler and Hutter 2013). Also, it is stated in the literature that on one hand Xenophobia revolves around the aspect of emotions and sentiments (Reynolds and Vine 1987), but on the other hand it should be studied as an activity, as a violent practice because it is rooted in the symbolic violence of everyday life (Bronwyn 2002). In this way events depicting physical violence are regarded legitimate for measuring Xenophobia. Moreover, the use of event extraction for predictive analysis is a very challenging and far from trivial task, whose potential justifies its impact within the literature (Boschee, Natarajan and Weischedel 2013).

All the above-mentioned projects implement methodologies varying from completely unsupervised, purely data-driven approaches, to knowledge-driven methods based on domain experts. Hybrid frameworks have also been used (Hogenboom et al. 2011). Our approach is a hybrid method which involves human-in-the-loop.

3. Extracting and Analyzing Events

The framework that was designed and implemented for the detection of Events within news data, is data driven and consists of five distinct phases:



Codebook

Data Collection

Data Exploration

Event Analysis

Data Visualization

Figure 1: Event Extraction Methodology

- First, political and social scientists in collaboration with computational scientists developed a **Codebook**, which depicts the taxonomy for the events under research along with their structural components and the intended attributes.

- In the **Data Collection** phase, data were collected from seven different news sources, both printed and electronic agencies. The data were appropriately prepared and stored in the repository. Data preparation included tackling data normalization issues and transforming them to a human readable corpus.
- During **Data Exploration**, human experts that helped develop the codebook were also involved. They explored the collected data, providing valuable insights. This was an iterative procedure, as simultaneously the Codebook was modified and improved, until it was finalized according to the exploration. The outcome of this phase was event oriented data collections.
- **Event Analysis** was conducted following an information extraction paradigm. The data were modelled per specific information types and a **Knowledge Base** was extracted. It comprises several Text Analytics pipelines, which detect all the Event constituents and the links among them. Firstly, an NLP suite performs pre-processing over raw text, producing annotations for Tokens, Lemmas, Chunks, Syntactic relations and Named Entities. This output is given as input to the Event Analysis System, a cascade of Finite State Transducers (FST) which identifies all the different information types and associates them to form an Event tuple. The result is a Knowledge Base that populates the final Database.
- Finally, **Data Visualization** is an important phase of the research cycle. During this stage, the results of the Information Extraction are visualized in various ways, making them explorable, comprehensible and thus more easily interpretable.

3.1 Codebook

The development of the Event Database included two milestones. The first one was the design of the coding schema and the second one its computational implementation. In the first phase, political and social scientists in association with computational scientists, designed and described in details the coding framework for the events related to the scope of the project, namely Xenophobia in Greece.

3.1.1 General Principles

This Codebook was designed following specific principles.

- One of the most important steps was the selection of the event types that are considered relevant to the subject of the project and which the research team believed that would assist with the interpretation of the research goal, namely Xenophobia in Greece. To that direction, an event Taxonomy was constructed comprising two major Event Categories, to wit Physical Attacks and Protests along with several event types for each category.
- Each event type was then clearly defined by the political scientists' team and communicated to the computational team to ensure that a common sense is built.
- It has also been crucial to describe in an unambiguous and comprehensive manner all the elements that constitute the event tuple, namely the entities involved as Actors or Targets, and elements defining the event such as the Location and the Time it took place.
- Representative examples were given for each event type and event constituent, illuminating the event tuple's structure.
- Finally, the expected attributes of the entities involved in the events were defined and described in details. Such attributes include for example defining the nationality of each detected entity.

3.1.2 The structure of the Event

The coding unit of the analysis is the event. Two major categories of events as well as various subcategories or event types are considered according to the scope of the project. In particular, this study examines two major event categories, namely Physical Attacks and Protests, each of which breaks down into distinct event types, which will be described hereafter. An event comprises a tuple containing several types of information. More specifically, an event consists of six different elements, and every element is attributed certain features, as detailed below in each respective section. It needs to be noted that there is a slight difference between the two event categories as far as the constituents of the tuple are concerned. Five out of six are common and the last one differs in the way described further on. Hence, the components of the event tuples are:

1. The event, namely the word or phrase representing an event type under examination, which is located within the text (**EVENT**). Features: *Event type*.
2. The actor, meaning the entity that performs each event instance (**ACTOR**). Features: *Summary*, *Target Group*, *Nationality*, *Age*, *Sex*, *Status*.
3. The target, namely the entity to whom the action is addressed (**TARGET**). Features: *Summary*, *Target Group*, *Nationality*, *Age*, *Sex*, *Status*.
4. The location where the event took place (**LOCATION**). Features: *Category*.
5. The time at which the event happened (**TIME**). Features: *Day*, *Month*, *Year*.
6. The confidence element, only for Physical Attacks (**CONFIDENCE**). This type of information captures whether in the article there is any indication that an Actor of an assault may not be the actual perpetrator. Features: *Degree*.
7. The Issue of a Protest Event (**ISSUE**), i.e. what the protest is about. Features: *Category*.

Therefore, each record in the Event Database comprises of the six aforementioned constituents. Ideally, all these elements in the tuple are completed, though this is not always the case. It is possible and quite common that some of them are missing. The component that is central is the Event. Only if an event lexicalization has been detected, can a candidate entity become Actor or Target, otherwise it remains a candidate. Moreover, for an event tuple to be considered as such and consequently to be recorded in the database, apart from the Event element, at least one of the {Actor, Target, Issue} is obligatory.

Some illustrative examples can be seen below.

ACTOR	CONFIDENCE	EVENT	TARGET	LOCATION	TIME
A 24-year-old American	*is accused of involvement*	*in the arson*	*of the Synagogue*	*at Chania*	*on Thursday.*

ACTOR	EVENT	TIME	TARGET	LOCATION	ISSUE
Employees of the General Secretariat for Research and Technology	*held a* ***protest rally***	*yesterday*	*outside Parliament*		*opposing the passage of the bill for research and technology.*

3.1.3 Event Coding Schema

3.1.3.a Physical Attacks

In this section, the event types that fall into the broader category "Physical Attacks" are coded. First, an index of all the event types is presented, followed by each type's definition and illustrated through an example.

Index

01. Assault

    011. Assault against life

    012. Violent Assault

    013. Sexual Assault

    014. Verbal Attack

02. Attack against Property

021. Attack against Personal Property

022. Attack against Religious Property

01. **Assault**: use of any form of violence against individuals, other than the ones specified below.

Ex. *Konstantinos Kontomou's four racist **attacks** against Pakistanis in the region of Metamorphosis in September 2012*.

011. **Assault against life**: physical assault to individuals aiming to kill them, assassination or attempt to assassinate.

Ex. *Unknown gunmen **shot** against policemen during a routine check, this morning at 7.30 a.m. at Ilissia*.

012. **Violent Assault**: attack physically to individuals without the intention to kill. Torturing and kidnapping also fall within this event type.

Ex. *The Coast Guard's chief imposed sanctions on six port officials that are involved in the **abuse** of three immigrants at Souda in June 2001*.

*It's about the appalling **abuse** of a 15-year-old girl by a young Pakistani in Paros*.

013. **Sexual Assault**: use force for sexual actions, sexually abuse.

Ex. *A 47-year-old teacher is held in prison for purportedly **molesting** seven of his students in the school he was teaching the last few years, in a village near Serres*.

014. **Verbal Attack**: verbally abuse, insult. This event type also includes threats.

Ex. *Serious incidents occurred on Wednesday night at Chania, when Golden Dawn's supporters* ***hurled insults*** *at migrants*.

02. **Attack against property**: cause damage or destroy movable or immovable property.

Ex. *Earlier, on 01:15, unidentified persons* ***threw Molotov cocktails*** *at the Ministry of Culture, on Mpoumpoulinas Street, thereby causing damage to the entrance of the building*.

021. **Attack against Personal Property**: damage or destroy property belonging to individuals.

Ex. *In the same evening, in Markoni region at Elaionas, unidentified individuals* ***attacked*** *at Pakistanis' house*.

022. **Attack against Religious Property**: damage, arson or destruction of places of worship, collective monuments or cemeteries.

Ex. *A 24-year-old American accused of involvement in the* ***arson*** *of the Synagogue at Chania, was released on bail on Thursday afternoon*.

3.1.3.b Protests

In this section, the event types which are classified as “Protests” are explained. Firstly, all of them are indexed and then each is defined and clarified through an example.

Index

- 01. Protest
    - 011. Demonstration/March
        - 0111. Motorized March
        - 0112. Sit Down
    - 012. Strike
    - 013. Hunger Strike
    - 014. Blockade
    - 015. Occupation
    - 016. Signature Collection
    - 017. Boycott
    - 018. Symbolic Violence
    - 019. Revolt
    - 020. Violent Demonstration

01. **Protest**: this broad event type covers every protest expressed as a collective or individual action.

Ex. *On Monday morning in Nea Vissa's square, residents of the region* ***protested*** *against the construction of the fence in Evros and the Operational Border Surveillance Centre*.

011. **Demonstration/March**: express collective discontent, gather and/or rally.

Ex. *Employees of the General Secretariat for Research and Technology held a* ***protest rally*** *yesterday outside Parliament, opposing the passage of the bill for research and technology*.

0111. **Motorized March**: rally using motor vehicles.

Ex. *Local authority workers continue their fight against the suspensions and redundancies the three-party government is attempting to impose. Today, the POE-OTA* ***organizes motorized protest marches*** *in every prefectural capital*.

0112. **Sit Down**: peacefully protest by sitting, usually on a public area.

Ex. *Workers on work experience (trainee) programs marched yesterday from Omonoia to Syntagma and* ***staged a sit-in protest*** *outside the Parliament, repeating their demands to be instated as civil servants*.

012. **Strike**: refuse to work as a protest, expressing certain demands.

Ex. *Employees at the movement of goods at the premises of Cosco, in Piraeus,* ***are on a strike*** *since this morning, protesting about working conditions*.

013. **Hunger Strike**: refuse to receive food as an act of protest.

Ex. *15 Arab economic migrants* ***went on hunger strike*** *yesterday, demanding Greek work permits.*

014. **Blockade**: obstruct passage, close off roads, railways.

Ex. *Yesterday on Crete, producers* ***blocked*** *both the old and the new highway at the Heraklion airport junction, demanding that their agricultural products be transported immediately*.

015. **Occupation**: occupy an area or building.

Ex. *Employees at EYATH (Thessaloniki Water Supply & Sewerage) continue for second day the* ***occupation*** *of the company's central offices demanding the signing of a new collective labor agreement*.

016. **Signature Collection**: gather signatures to promote specific claims.

Ex. *Athens University of Economics started* ***collecting signatures*** *against the abolition of its autonomy*.

017. **Boycott**: voluntarily abstain from buying or consuming products of a certain company or country.

Ex. *Lawyers, engineers and doctors have voted to jointly mobilize against measures imposed by the Troika and to* ***boycott*** *German products*.

018. **Symbolic Violence**: actions with symbolic connotations, such as puppet or flag burning, egg or yogurt throwing.

Ex. *Shortly afterwards, members of Pontian Greeks' Associations* ***burnt*** *a Turkish flag, while offensive slogans were addressed to politicians*.

019. **Revolt**: riot as a collective action, mainly referring to prisoners.

Ex. *Yesterday, in Korydallos prison, immigrants that are still imprisoned only because they cannot be expelled, held a two-hours* ***riot*** *demanding their trials to be held sooner*.

020. **Violent Demonstration**: rally or demonstrate forcefully, protests that turn violent, causing damages.

Ex. *The anti-war demonstration was organized by the Greek Social Forum, but the mobilization was marred yesterday by* ***violent incidents*** *between demonstrators and the police*.

3.1.4 Actor Coding

As Actor, the perpetrator of a Physical Attack or the protester at a Protest Event is considered. Several entities can be attributed the label Actor, varying from individuals to companies and organizations or whole countries. Thus, various tools are used for detecting entities and assigning them a Candidate label at first. In a second stage, only the entities that are linked to a specific event are characterized as Actors.

Apart from the entity itself, there are other features of the Actors that are required by the scope of the present project. Since the aim is to examine the phenomenon of xenophobia, the notion of what is considered as foreign or strange needs to be defined. Therefore, the sense of foreign has been restricted to immigrants and refugees living in Greece, while social and political scientists have determined concrete Target Groups (see section 3.1.4.2 below). Additionally, the implemented system intends to capture demographics such as Age and Sex.

Several dictionaries along with computational tools exploiting contextual and semantic information have been developed aiming at detecting Actors along with their aforesaid features.

In the following sections, all the features which are attributed to Actors are described in detail.

3.1.4.1 Actor Summary

The lexicalization of an entity – individual, organization, group – that refer to the Actor of an event, is the first element detected in the news article. Based on semantic rules, these entities are systematically classified into the following specific categories.

- Anarchists and leftist organizations and groups
- Central banks
- Churches and Religious organizations and groups
- Consumers' organizations and groups
- Economists and financial experts
- Education professionals
- Educational organizations
- Employees
- Employees in municipalities and regions
- Employer organizations
- Environmental organizations and groups
- Ethnic minorities
- Farmers
- Firms
- Freelancers working on transport
- Friend

- Government executives
- Health sector employees
- Higher education and research institutions
- Judiciary
- Legislative and parliament members
- Mass transport employees
- Media and journalists
- Migrants and asylum seekers
- Municipal authorities
- NGOs
- Pensioners/Organizations and groups for the elderly
- Other
- Police, military and security forces
- Political parties and organizations
- Prisoners
- Racist and extreme right organizations and groups
- Refugees
- Relative
- Solidarity and human rights organizations and groups
- State owned enterprises
- Students and pupils
- Terrorist groups and rebel forces
- Tertiary Trade Unions
- The general public
- Trade Unions
- Unknown
- Whole countries
- Women's organizations and groups

3.1.4.2 Actor Target Group

Eight Target Groups were initially defined by political and social scientists, considering the population in Greece, the migration waves that fall into the period this project examines and the well-established prejudices that are embedded in the Greeks concerning specific national groups. During the project's implementation, 2 more Target Groups were added, covering the current affairs concerning the refugee crisis that arose. The final 10 Target Groups examined are recorded below.

- TG0: Refugees.
- TG1: Pakistani.
- TG2: Albanians.
- TG3: Rumanians.
- TG4: Syrians.
- TG5: Muslims.
- TG6: Jews.
- TG7: Germans.
- TG8: Roma.
- TG9: Immigrants.

It should be noted, first that all the other nationalities are grouped as Other and second that, apart from the Target Groups, there is also the Control Group against which the comparisons are made, namely Greeks.

3.1.4.3 Actor Nationality

Under this attribute, the lexicalization of each Actor's ethnic identity is recorded as found within the text. In fact, this is an auxiliary element to determine the Target Group of the Actor.

3.1.4.4 Actor Age

This feature records the age of the Actor as it is referred within the text, usually as a two-digits number but also lexicalized, as a word. This attribute was considered significant and included in the coding schema, aiming to capture demographic differentiations between Actors.

3.1.4.5 Actor Sex

This is a two-values string variable, namely Male and Female. It detects the sex of the Actor and documents it. The working team of political and social scientists considered this feature important in the context of demographics too.

3.1.4.6 Actor Status

This attribute refers to the legal status of immigration, i.e. it can be filled in with two string values, viz. legal or illegal. In all the rest cases, it is left blank.

3.1.5 Target Coding

The Target entities of the events extracted and stored in the Event Database, have the same attributes as the Actor entities.

3.1.6 Location Coding

Each event is attempted to be linked to the particular Location where it took place. This type of information captures the lexicalization of the event location. The goal is to translate each location to its coordinates, latitude and longitude, so as to be able to map them. However, this turned out to be a non-trivial task, considering the vast amount of data processed. Thus, the analysis team decided to prioritize the other elements of the events and postpone the geolocation for a later stage of research.

3.1.6.1 Location Category

Given the fact that this project examines Xenophobia in Greece, it was crucial to discriminate between events happening in Greece or anywhere else. To this end, each location name was categorized according to the country it is belongs to. In this way, the following categories occurred.

- Continent
- Countries Africa
- Countries Asia
- Countries Caribbean
- Countries Central America
- Countries North America
- Countries Oceania
- Countries South America
- Desert India
- Desert Namibia
- European capital
- European Countries
- Greece
- Mountain foreign
- Peninsula
- Province France
- Region Europe
- Region Germany
- Region Middle East
- Region Serbia
- Region South America
- River foreign
- State USA

- Town Adjerbaijan
- Town Afghanistan
- Town Albania
- Town Andorra
- Town Armenia
- Town Australia
- Town Austria
- Town Belarus
- Town Belgium
- Town Bosnia Herzegovina
- Town Brazil
- Town Bulgaria
- Town China
- Town Crimea
- Town Croatia
- Town Cyprus
- Town Czech Republic
- Town Denmark
- Town Egypt
- Town Estonia
- Town Finland
- Town France
- Town FYROM
- Town Georgia
- Town Germany
- Town Holland
- Town Hungary
- Town Iceland
- Town India
- Town India
- Town Iraq

- Town Ireland
- Town Israel
- Town Italy
- Town Japan
- Town Latvia
- Town Lichtenstein
- Town Lithuania
- Town Luxemburg
- Town Malta
- Town Moldova
- Town Montenegro
- Town Nepal
- Town New Zealand
- Town North America
- Town Norway
- Town Oceania
- Town Pakistan
- Town Peru
- Town Philippines
- Town Poland
- Town Portugal
- Town Romania
- Town Russia
- Town Samoa
- Town Saudi Arabia
- Town Scotland
- Town Serbia
- Town Slovakia
- Town Slovenia
- Town Somalia
- Town South Africa

- Town Spain
- Town Sweden
- Town Switzerland
- Town Syria
- Town Thailand
- Town Turkey
- Town UAE
- Town UK
- Town Ukraine
- Town USA
- Town Wales

3.1.7 Time Coding

Temporal location of events is an unresolved issue in the field of information extraction. Although several efforts have been made to that direction and various techniques have been proposed, it remains a demanding, non-trivial task for automated computational systems. In the context of this project, the temporal expressions referred in the text are recorded and the intention was to map them to actual dates. Though, since this is a strenuous and time-consuming procedure and considering all the other types of information the developed system needed to capture, the day of the article's publication was decided to be recorded as the date of the event. This decision was based on the fact that in the clear majority of events, the day of the newspaper issue or the previous one is the actual date an event took place. Moreover, in some cases the article does not report the exact time of the event. Finally, the analysis is conducted in annual or monthly intervals, therefore defining the exact day is not a prerequisite. Consequently, exact time classification was not addressed at this point – it remains a goal of our lab for the future -, given that no time sequence was needed. Also, we were interested in classifying the events into monthly intervals so there is not deviation in this time window.

Given this background, the date of each event is recorded in the Event Database using three variables, namely Year, Month and Day, allowing for analyses of specific time periods.

3.1.8 Confidence Coding

As indicated above, this information type is relevant only to the Physical Attack events. The scope is to capture the certainty degree to which an Actor of an attack is actually the one referred in the text. This type of information was considered important due to the way such events are reported in the news. In other words, news agencies are considered to be objective and just report facts. Of course, this is the case only for physical attacks where there is the possibility of somebody being falsely accused. In the case of Protest events on the other hand, the Actor is unambiguous and thus the Confidence element is not necessary.

3.1.8.1 Confidence Degree

This is a two-values variable, with True and False being the possible values. When the variable is set to True, it means that the Actor attributed to an event is certainly the entity that did it. In contrary, when the value is False, it signifies that the Actor may not be the actual perpetrator. This is due to the way mainly criminal acts are presented in the press. There are some restrictions posed by the Justice, concerning people that have not yet been convicted. For example, in the sentence*: "A 23-years-old Greek is accused of having killed his 20-years-old colleague"*, we are not completely sure that the 23-years-old Greek is actually the one who committed the murder, so the Confidence Degree value is set to False.

3.1.9 Issue Coding

As mentioned before, the Issue element is coded only for Protest Events. It is a string variable, capturing the word, phrase or structure depicting the subject matter of a protest, or else, the reason why people is protesting.

3.1.9.1 Issue Category

The issue of a protest event as recorded within a text, is then categorized according to the topic it is related to. These topic categories are the following:

- Agricultural Affairs
- Anti-War
- Budget and Economic Affairs
- Education and Culture
- Employment and Social Affairs
- Environment, Energy and Waste Management
- European Union
- Government Policies
- Health
- Human and Civil Rights
- Justice and Legislation
- Local Governance
- Media
- Immigration
- Other
- Police and Other Violence
- Political System
- Racism and Fascism
- Transport

3.2 Data Collection

For the Event Extraction task, a large collection of news data from different sources both printed and electronic were used. The articles are in Greek and metadata (section labels, headlines, names of authors) were gathered for each along with the text itself. More specifically, the data were collected from 7 news agencies in Greece, as described below:

1. **Avgi**: 792.715 articles from the printed version of the newspaper for the time period 1996-2015.
2. **Kathimerini**: 282.621 articles of the electronic version of the newspaper for the time period 2002-2006 and 2009-2012.
3. **Eleftherotypia**: 429.364 articles of the electronic version of the newspaper. Time span 2002-2006 and 2008-2014.
4. **Rizospastis**: 725.108 articles of the printed version of the newspaper. Time period 1995-07/2016.
5. **TaNea**: 330.190 articles from the electronic edition of the newspaper for the time period 1997-2007.
6. **In.gr:** 428.880 articles from this media website for the period 1999-21/09/2016 (date of data retrieval).
7. **Naftemporiki:** 649.259 articles of the online version for the period 2000-21/09/2016 (date when the data were retrieved).

Hence, in total 3.638.137 news articles for a time span of more than 20 years, specifically 1995-2016 were gathered, prepared and stored. Data preparation included tackling data normalization problems and transforming the data to a human readable corpus.

3.3 Data Exploration

The phase of data exploration was vital to the analysis, since the followed approach is data-driven and sets out to incorporate human-in-the-loop. Therefore, as the first step of Data Analysis, human experts explored the collected datasets using queries. The aim of this procedure was to determine the ways in which each event type and its constituents, are expressed and lexicalized. The queries started as simple word or phrase queries and resulted to complex ones using Boolean operators.

The exploration stage was also crucial for filtering the collected bulk of data and group them into event-oriented data clusters. This procedure was interactive and followed several iterations, as it was directed by the Codebook (see Section 3.1) and the Codebook was altered and enriched in line with the results of exploration.

The main goal of the Explorative Analysis was to better understand and obtain a wide view of the whole dataset. Given that the available datasets came from various media sources, probably reflecting ideological and idiosyncratic characteristics, it was essential to examine the different ways and linguistic means that each news agency uses to report the same event. To this end, a full text search application was developed. Exploiting the ELK stack (Elastic, Logstassh, Kibana), PALOMAR, a platform for automated and scalable data processing was developed and used to index the data and make the datasets available to the users. The core functionalities of the interface included the ability for the user to make full-text queries, simple or compound, select articles, inspect them and save the search. They are also able to come back to the queries and modify them. Subsequently, in the data analysis, the saved queries along with the articles indicated as relevant were retrieved and stored into data clusters, one for each event type.

3.4 Event Analysis

Event Analysis is a multifaceted task. The overall event extraction framework implemented in the context of this project is data-driven yet linguistically oriented. Its foundations lay on political and social sciences, additionally incorporating human-in-the-loop. The adopted approach is to first detect each structural component comprising the analytics stack and afterwards to relate the right elements and create event tuples, which are finally recorded into a Knowledge Base. The methodology employed is semi-supervised, in the sense that a small fraction of data was labeled and used for the development of the system. Moreover, it is linguistically driven, thus morphosyntactic information from basic NLP tools is used to approach the information types comprising the event tuple as it is defined in the Codebook.

The general rationale for extracting events is bootstrapping (Papanikolaou et al., 2016), in the sense that every module builds over the annotations produced by previous modules (Papageorgiou and Papanikolaou, 2016). At the first stage, a Natural Language Processing tools suite is used for performing pre-processing over raw text and producing annotations for Tokens, Lemmas, Chunks, Syntactic relations and Named Entities. In the next phase, the pre-processing output is given as input to the Event Analysis Unit, which performs two subsequent tasks. First, the elements comprising an Event are detected and subsequently linguistic rules based on shallow syntactic relations link the correct components to form an event tuple, which is then recorded into the Event Database. The Event Analysis system is implemented as Finite State Transducers (FSTs) using Gate JAPE patterns[1]. These FSTs process annotation streams utilizing regular expressions to create generalized rules. Moreover, they are ordered in a cascade, so that the output of an FST is given as input to the next transducer. Figure 2 depicts the Data Analytics stack for Event Extraction.

[1] https://gate.ac.uk/

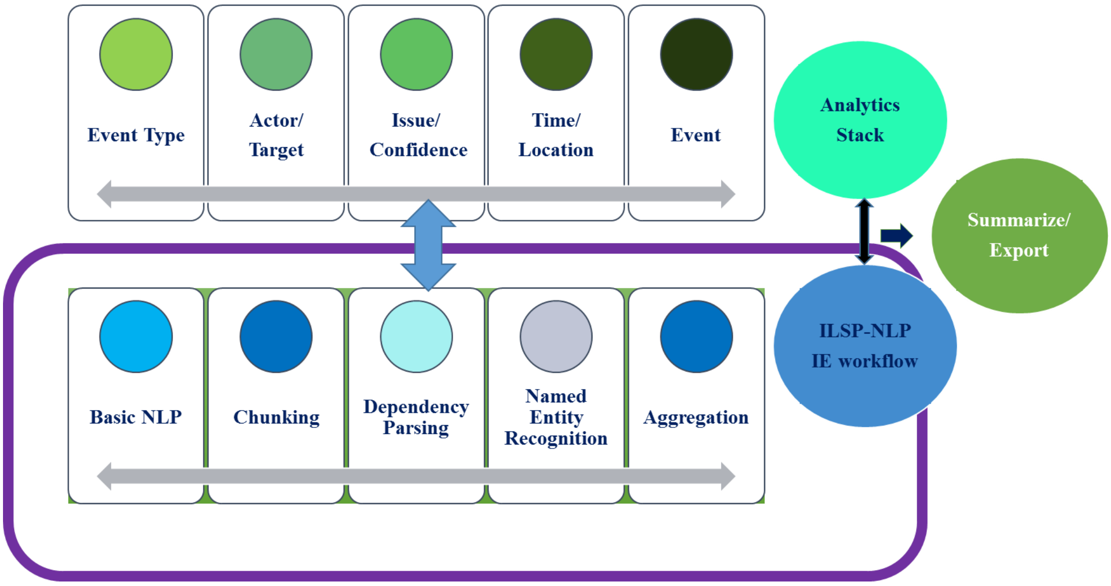


Figure 2: The Event Analytics stack

The above presented NLP workflow is fed with textual content. At the first phase, it performs Tokenizing, POS Tagging, Lemmatizing, Chunking and Dependency Parsing. After that, the Named Entity Recognizer and Classifier (NERC) module, recognizes and classifies Named Entities into four major categories, i.e. Persons, Organizations, Locations and Facilities. The output of this processing is then forwarded to the Event Analysis Unit, where Nominals – viz nominally expressed entities – are detected and labeled as Candidate entities together with Person and Organization annotations. In the next step, Time expressions are identified and the Issue information type is annotated, while another module caters for the identification of each Event type. The following stage is to determine whether a Candidate entity can be assigned as Actor or Target to a specific Event. At the final step, the detected and linked annotations comprising event tuples are extracted and recorded into the Event Database. The following sections present more details about each phase of the workflow.

3.4.1 Pre-processing

Pre-processing is an integral part of the implemented methodology. During this phase, basic NLP annotations are produced, upon which the Event Analysis Unit builds to create the final output, namely the Event Database. Figure 3 depicts the general architecture and outlines the tools used for this purpose.

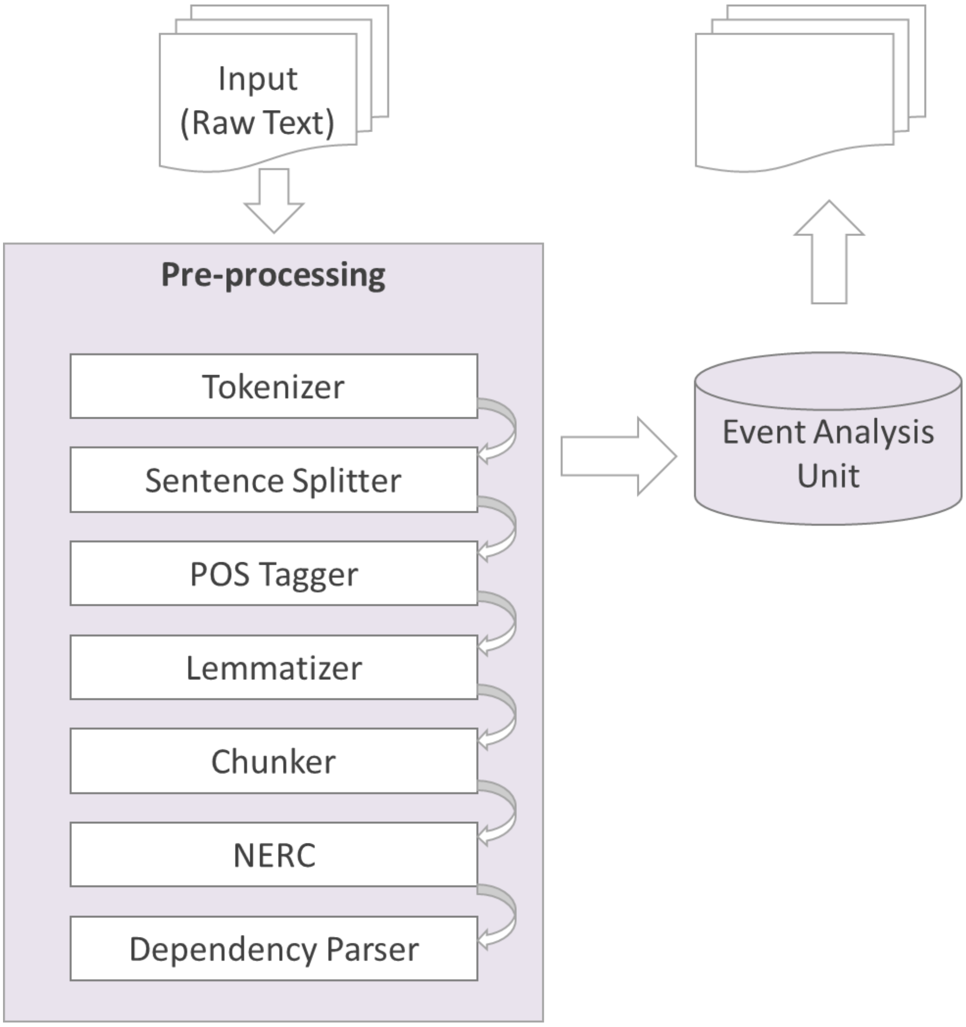


Figure 3: Workflow for Event Extraction

This phase, as is presented above, involves Text Analytics pipelines which process the documents stored in the repository with workflows that include:

- Segmentation, i.e. recognizing paragraph, sentence and token boundaries. A token can be a word, a number or even punctuation marks.
- Part of speech tagging, i.e. assigning morphosyntactic categories and features to each token. For example, the token *strike* can be a Noun or a Verb depending on the context.
- Lemmatization, i.e. retrieving the base form from an inflected token. For instance, both *strike* and *strikes* are attributed to the lemma *strike*.

- Chunking, i.e. performing a shallow syntactic parsing, discovering clauses and syntactic constituents such as Noun and Verb Phrases.
- Entity recognition, i.e. recognizing Named Entities and classifying them into four categories: *Person*, *Organization*, *Location*, *Facility*.
- Parsing, i.e. determining the syntactic structure of a sentence.

Each of these tasks is performed using tools designed and developed specifically for the Greek language by the Institute for Language and Speech Processing. These tools are fully integrated and curated in the clarin:el infrastructure.

3.4.2 Event Analysis Unit

The Event Analysis Unit framework is based on linguistics, in the sense that semantics and syntactic parsing patterns are used. It consists of several modules which seek to detect the structural components of the Event and create links among them. Figures 4 and 5 below depicts the general architecture of the system for Physical Attacks and Protest Events detection.

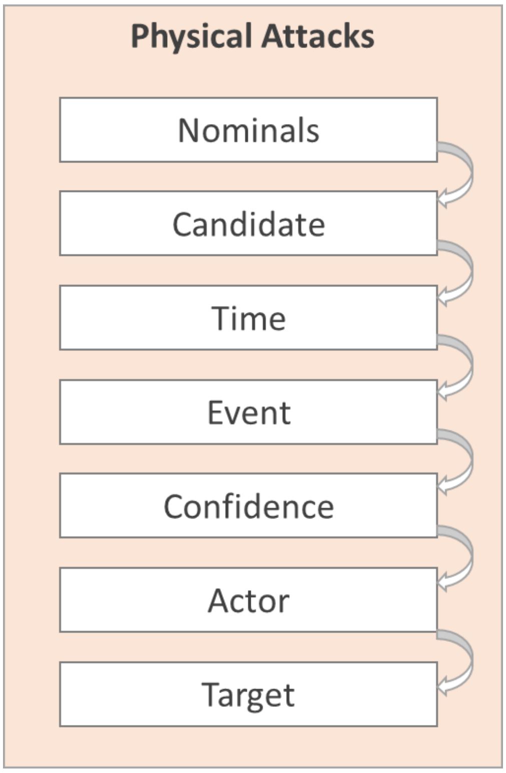


Figure 4: Event Unit Architecture (Physical Attacks)

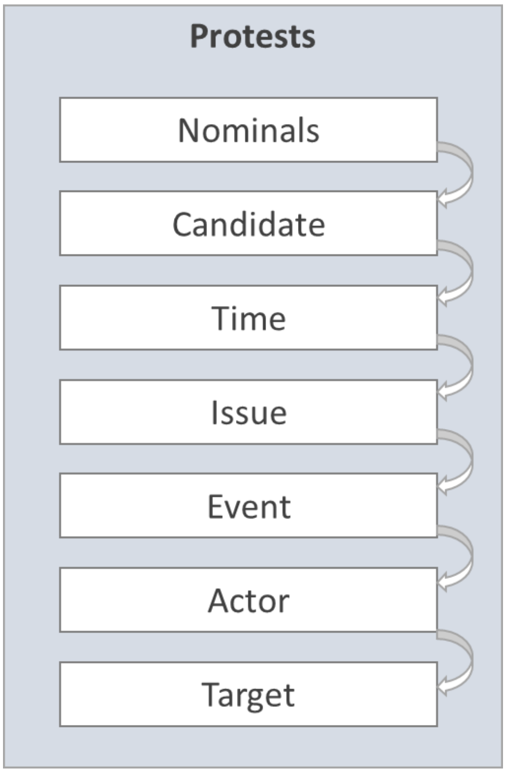


Figure 5: Event Unit Architecture (Protests)

Every phase of the implemented methodology, as is shown in the above figures, along with the developed tools, are described in details hereafter. Illustrative examples of events are used to facilitate the comprehension of the system and its architecture.

1. *15 Arab economic migrants went on hunger strike yesterday, demanding Greek work permits*.
2. *Athens University of Economics started collecting signatures against the abolition of its autonomy*.
3. *On 28 March 2013, in the center of Xanthi, a 30-year-old Greek Muslim suffered a stub wound by members of Golden Dawn*.
4. *Deputy Minister for Defence Fofi Gennimata was verbally abused by workers in the defence industry*.
5. *A group of 30 people attacked last Tuesday an immigrants' house in the city of Kalamata*.
6. *A 47-year-old teacher is held in prison for purportedly molesting seven of his students in the school he was teaching the last few years, in a village near Serres*.
7. *Panhellenic Seamen's Federation are on a 48-hour strike on Holy Tuesday and Wednesday, 10 and 11 of April, demanding a solution to the problem of seamen's unemployment, the signature of collective working agreements for 2012 and the end of the pensions' reduction*.
8. *Two officers of the riot police are accused of beating two Albanian workers who filed lawsuits against them*.

As aforementioned, the first module in the pipeline is the Nominals module, which utilizes POS tags, Lemmas and Chunks to identify entities that are not named, but some nominal expression referring to them is present in the text. This is the case, for example, when a person is mentioned by his/her profession, religion, nationality or other property rather than by name, e.g. "*the Prime Minister*" instead of "*Alexis Tsipras*". This tool's objective is to annotate the most informative reference. Hence, in example (1) above, not only *migrants* but *15 Arab economic migrants* is annotated as Nominal, providing information about the number, the nationality and the status of the agents. Similarly, in example (3), *30-year-old Greek Muslim* and *members of*

Golden Dawn are recognized. Moreover, this module attributes specific features to the detected entities (see Section 3.1.4). For instance, based on semantics, they are grouped in categories as indicated by the Codebook (see Section 3.1.4.1). Additionally, the Age and Status, if mentioned, are recorded, as well as the Nationality, according to which entities are classified into Target Groups.

At the next step, all the entities that have been annotated as Nominals, together with the Person and Organization annotations from the NERC, are assigned the label Candidate, which means that they all are potential Actors or Targets of Events. For example, *Athens University of Economics* in example (2), identified by the NERC as Organization, is renamed to Candidate while it inherits all the features attributed to it, namely that its summary is *Educational Organization*. Nevertheless, this module's goal is not to simply rename annotations; it also consolidates them into more coherent lexical units. For instance, in example (4) above, the NERC would have identified *Fofi Gennimata* as Person and the Nominals module would have recognized *Deputy Minister for Defence,* designating a person's Title or the rank within a government board. The Candidate module merges these two annotations into one which is more complete and informative, namely *Deputy Minister for Defence Fofi Gennimata*.

Afterwards, time expressions are located, in a two-steps procedure. First, using a computational lexicon build for the purposes of the project, temporals (months, seasons e.tc.) are detected. Then, a set of linguistic rules built upon the time lexicon entries, are executed to discover more extended temporal phrases. For example, *last Tuesday* (example 5) is annotated by this module. However, as stated before (see Section 3.1.7), in the Event Database, though this information is recorded, Time information is normalized to the day the specific article was issued.

The next element to be located is the Issue which, as already indicated, concerns only the Protest Events. This information type strongly relies on semantics, hence a list of trigger words together with their syntactic complements are combined into patterns intending to identify *Issue*. In such a pattern the keyword "*demand*" is associated with its syntactic complement, namely a Noun Phrase in accusative. For instance, in example (1) the Issue is (*demanding) Greek work permits*. It should be noted that this module is designed to annotate up to three Issues, forming part of a coordinate structure. Thus, in example 7 *a solution to the problem of seamen's unemployment, the signature of collective working agreements for 2012* and *the end of the pensions' reduction* are recorded as Issue.

In the following phase, the Event information type is detected, both for Physical Attacks and Protests. It is important to note that both verbal and nominal event extraction is addressed, so these modules are activated only if at least one lexicalization, verbal or nominal, is discovered. Every such annotation is assigned certain features in terms of the semantics it conveys and the syntactic structure it requires. E.g. in example (5), *attacked* is annotated as Event and is assigned the value Attack for the attribute event type and the value VB5, indicating a verb that requires as object either a Noun Phrase or a Prepositional Phrase both in accusative. There are two ancillary lexica in this module, serving to distinguish between event types, in the context of Physical Attacks. The one, contains several real estate elements, divided into personal and religious property assets and is necessary for discriminating attacks against people from attacks against property. The second lexicon includes objects that can be used for assaults, such as knife or gun. This lexicon's entries are crucial in discriminating between several types of attack, e.g. attack against life or violent assault.

As far as the Physical Attacks are concerned, the module for detecting the Confidence information type is then activated. This is a computational lexicon comprising several words and phrases, which are topically related to the subject, namely violent events or crimes. For example, words like *confessed* or *was condemned* and phrases like *is accused of*, are located by this module, as in the sentence (8) *are accused of* is recorded as indicating Confidence.

Afterwards, a set of linguistic rules modelling shallow syntactic relations are triggered. These rules utilize the semantic information of Event combined with syntactic information so that *Candidate* entities can be assigned the label ***Actor*** or ***Target***, in accordance with their role. For example, in example (3) *30-year-old Greek Muslim* is recorded as Target while *members of Golden Dawn* as Actor. Finally, this cascading subsystem links all the detected components to create a tuple representing an *Event* at sentence level.

In the last phase, the processing window is expanded, spanning the sentence where an Event lexicalization was discovered, along with its anterior. Thus, if an Event annotation has not been successfully related to an Actor and/or a Target, this module attempts to assign these labels to a Candidate entity within the boundaries of the two sentences, using rules and constraints concerning its features. Such a constraint dictates that an entity categorized as *Government* cannot be the *Actor* of a *Strike*.

Subsequently, the output of the above expounded workflow, is a set of tuples, each depicting an Event with its structural components. For the aforementioned examples, the extracted tuples are the following:

1. <**Actor**: *15 Arab economic migrants*, **Event**: *went on hunger strike*, **Target**: NA, **Issue**: *Greek work permits*, **Time**: *yesterday*, **Location**: NA>
2. <**Actor**: *Athens University of Economics*, **Event**: *started collecting signatures*, **Target**: NA, **Issue**: *the abolition of its autonomy*, **Time**: NA, **Location**: NA>
3. <**Actor**: *members of Golden Dawn*, **Event**: *stub wound*, **Target**: *30-year-old Greek Muslim*, **Confidence**: NA, **Time**: *28 March 2013*, **Location**: *center of Xanthi*>
4. <**Actor**: *workers in the defence industry*, **Event**: *verbally abused*, **Target**: *Deputy Minister for Defence Fofi Gennimata*, **Confidence**: NA, **Time**: NA, **Location**: NA>
5. <**Actor**: *A group of 30 people*, **Event**: *attacked*, **Target**: *immigrants' house*, **Confidence**: NA, **Time**: *last Tuesday*, **Location**: *city of Kalamata* >
6. <**Actor**: *47-year-old teacher*, **Event**: *molesting*, **Target**: *seven of his students*, **Confidence**: *purportedly*, **Time**: NA, **Location**: *village near Serres* >
7. <**Actor**: *Panhellenic Seamen's Federation*, **Event**: *are on a 48-hour strike*, **Target**: NA, **Issue**: *a solution to the problem of seamen's unemployment, the signature of collective working agreements for 2012 and the end of the pensions'* reduction, **Time**: on *Holy Tuesday and Wednesday, 10 and 11 of April*, **Location**: NA>

8. <**Actor**: *Two officers of the riot police*, **Event**: *beating*, **Target**: *two Albanian workers*, **Confidence**: *are accused of*, **Time**: NA, **Location**: NA>

4. Results

The above presented methodology for Event Analysis resulted in the population of the Event Database. More specifically, 14 files were created in total, in a two-dimensions discrimination. On the one hand, the two major event categories, namely Physical Attacks and Protests were distinguished and on the other hand, one file for each of the seven data sources was produced. The reason for not aggregating the data into one file was the fact that the time coverage of the media sources was not exactly the same. Therefore, in order to be able to produce visualizations like timelines and other quantitative and qualitative analyses, it was considered appropriate to separate the data according to the source they resulted from. It needs to be noted that all the files containing the results have been uploaded and shared through the CLARIN infrastructure[2].

The table below depicts the records per data file:

FILE	RECORDS
PhysicalAttacks_Avgi	29.942
PhysicalAttacks_Eleftherotypia	18.943
PhysicalAttacks_InGr	21.196
PhysicalAttacks_Kathimerini	13.043
PhysicalAttacks_Naftemporiki	19.672
PhysicalAttacks_Nea	17.028
PhysicalAttacks_Rizospastis	34.990
ProtestEvents_Avgi	13.501
ProtestEvents_Eleftherotypia	6.553

[2] http://www.clarin.gr/

ProtestEvents_InGr	9.451
ProtestEvents_Kathimerini	4.167
ProtestEvents_Naftemporiki	8.451
ProtestEvents_Nea	4.091
ProtestEvents_Rizospastis	34.648

Table 1: Records per Event Data File

Each file containing data for events of Physical Attack, comprises 55 columns, each of which represents a specific information type expressed as a variable. For the files including Protest events the columns are 59. Most of the information types were described in the Codebook section (3.1), while there are also some columns with metadata. There are four types of variables, i.e. String, Numeric, Binary and Boolean variables, according to the type of information they capture. The content of each column is clarified below:

ID: numeric variable. A unique identification number for each event recorded in the Data base.

MEDIUM: string variable. The source from which the article came from, namely one of the {Avgi, Eleftherotypia, Ingr, Kathimerini, Naftemrporiki, Nea, Rizospastis}.

TITLE: string variable. The headline of the article.

ADAY: numeric variable (values: 1-31). The day of the newspaper issue.

AMONTH: numeric variable (values: 1-12). The month of the newspaper issue.

AYEAR: numeric variable (four digits). The year of the newspaper issue.

EDATE: string variable. The time expression related to the event as was found in the text (see Section 3.1.7).

LOCATION: string variable. The name of the location where an event took place, as reported in the article.

LOCCAT: string variable (specific values, see Section 3.1.6.1). The country of the event’s location.

ACT1, **ACT2**, **ACT3**: string variable. This variable records the first, second, third actor as it is reported in the text.

ACT1S, **ACT2S**, **ACT3S**: string variable (specific values, see Section 3.1.4.1). The category into which the first, second, third actor is clustered.

ACT1TG, **ACT2TG**, **ACT3TG**: numeric variable (values: 0-9). The Target Group of the first, second, third actor (see Section 3.1.4.2).

ACT1SEX, **ACT2SEX**, **ACT3SEX**: binary variable (values: Male, Female). The sex of the first, second, third actor.

ACT1AGE, **ACT2AGE**, **ACT3AGE**: numeric variable (values: 1-99). The age of the first, second, third actor as is recorded in the article.

ACT1NAT, **ACT2NAT**, **ACT3NAT**: string variable. The nationality of the first, second, third actor as documented in the text.

ACT1STATUS, **ACT2STATUS**, **ACT3STATUS**: binary variable (values: Legal, Illegal). The immigration status of the first, second, third actor.

EVENT: string variable. The event’s lexicalization as referred to in the article.

EVENTTYPE: string variable (predefined values, see Section 3.1.3). The type of the Physical Attack or Protest according to the events taxonomy defined in the Codebook.

TRG1, **TRG2**, **TRG3**: string variable. This variable records the first, second, third target as it is reported in the text.

TRG1S. **TRG2S**, **TRG3S**: string variable (specific values, see Section 3.1.5). The category into which the first, second, third target is clustered.

TRG1TG, **TRG2TG**, **TRG3TG**: numeric variable (values: 0-9). The Target Group of the first, second, third target (see Section 3.1.5).

TRG1SEX, **TRG2SEX**, **TRG3SEX**: binary variable (values: Male, Female). The sex of the first, second, third target.

TRG1AGE, **TRG2AGE**, **TRG3AGE**: numeric variable (values: 1-99). The age of the first, second, third target as is recorded in the article.

TRG1NAT, **TRG2NAT**, **TRG3NAT**: string variable. The nationality of the first, second, third target as documented in the text.

TRG1STATUS, **TRG2STATUS**, **TRG3STATUS**: binary variable (values: Legal, Illegal). The immigration status of the first, second, third target.

CONF: (applies only to Physical Attacks) string variable. The word or phrase that indicates the certainty or uncertainty of an actor (see Section 3.1.8).

CONFDEGREE: (applies only to Physical Attacks) Boolean variable. This variable is True if the text indicates an actor as certain and False if doubts are expressed about the actor.

ISSUE1, **ISSUE2**, **ISSUE3**: (applies only to Protests) string variable. The text excerpt that describes the first, second, third issue of a protest event.

ISSUE1CAT, **ISSUE2CAT**, **ISSUE3CAT**: (applies only to Protests) string variable (prespecified values, see Section 3.1.9). The topical category of the first, second, third issue.

5. Data Visualization

The Visualization phase is an integral part of the project as the results need to be visualized in different ways, making them understandable and easily usable for the human eye. That is crucial in order to be able to interpret them, find correlations or important insights and drive to conclusions according to the scope of the project.

In this context, several useful visualizations were produced from the results files. The great amount of information types that were extracted, allows for many different associations and graphs. Hence, the generated visualizations include charts, timelines, pies and word clouds. Moreover, there is the possibility to create more, filtering the results according to specific information types or attributes, configuring temporal windows or geolocating the results to produce information maps. Some of the most prominent visualizations used for answering the research questions of the project are presented below.

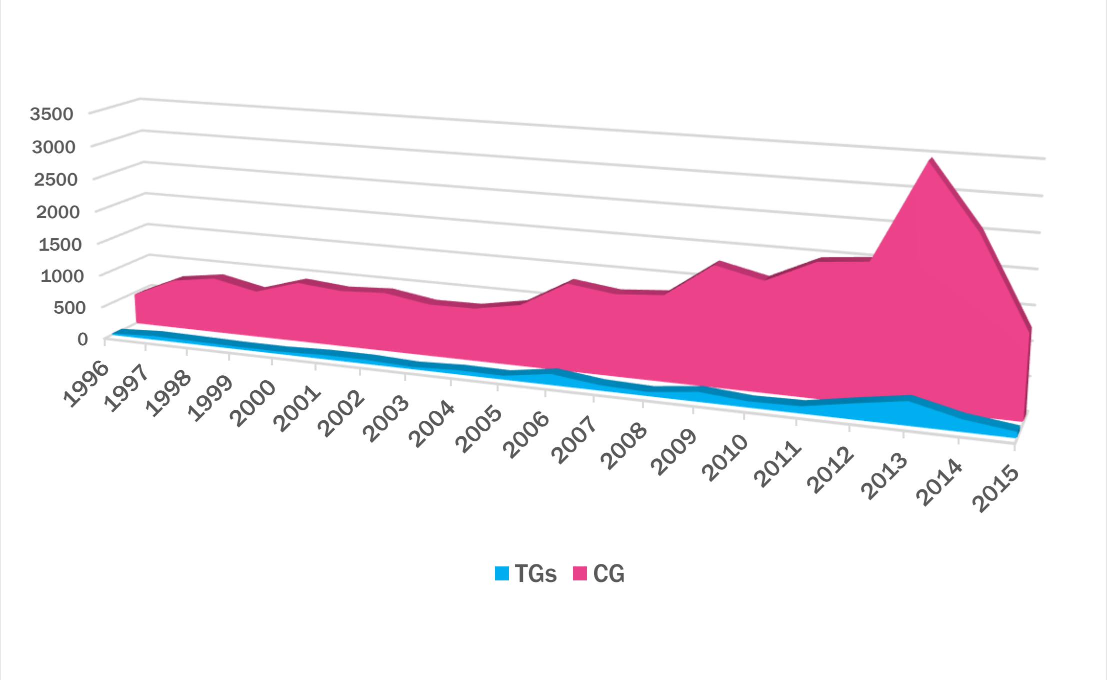


Figure 6: Physical Attacks against Target Groups compared to Control Group

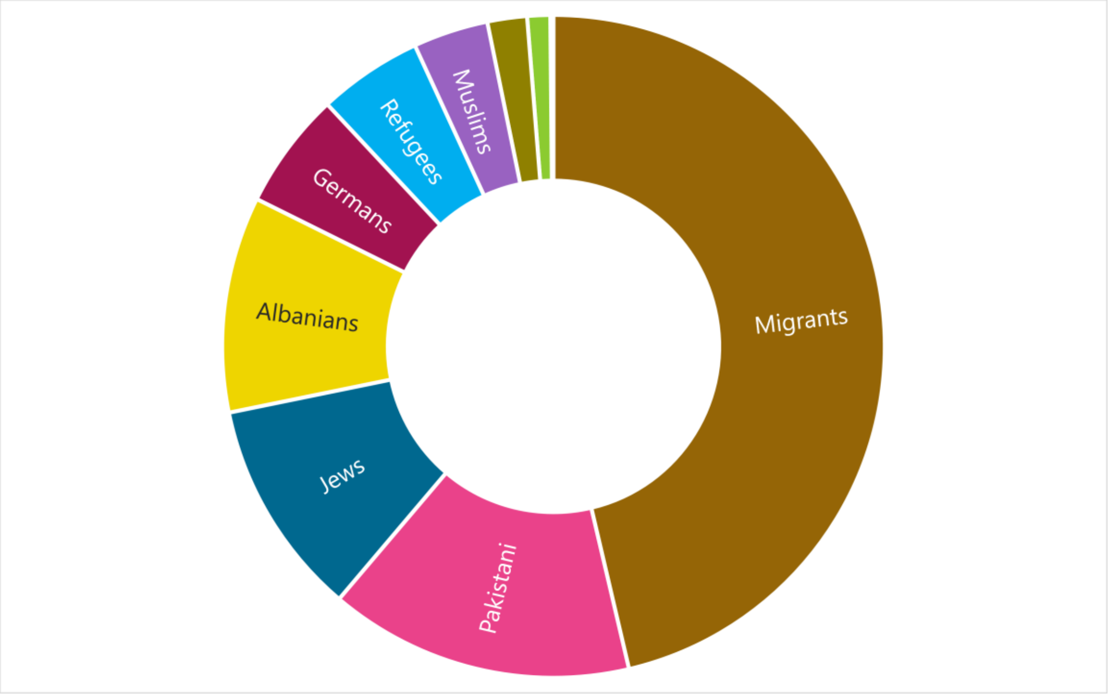


Figure 7: Mostly Attacked Target Groups

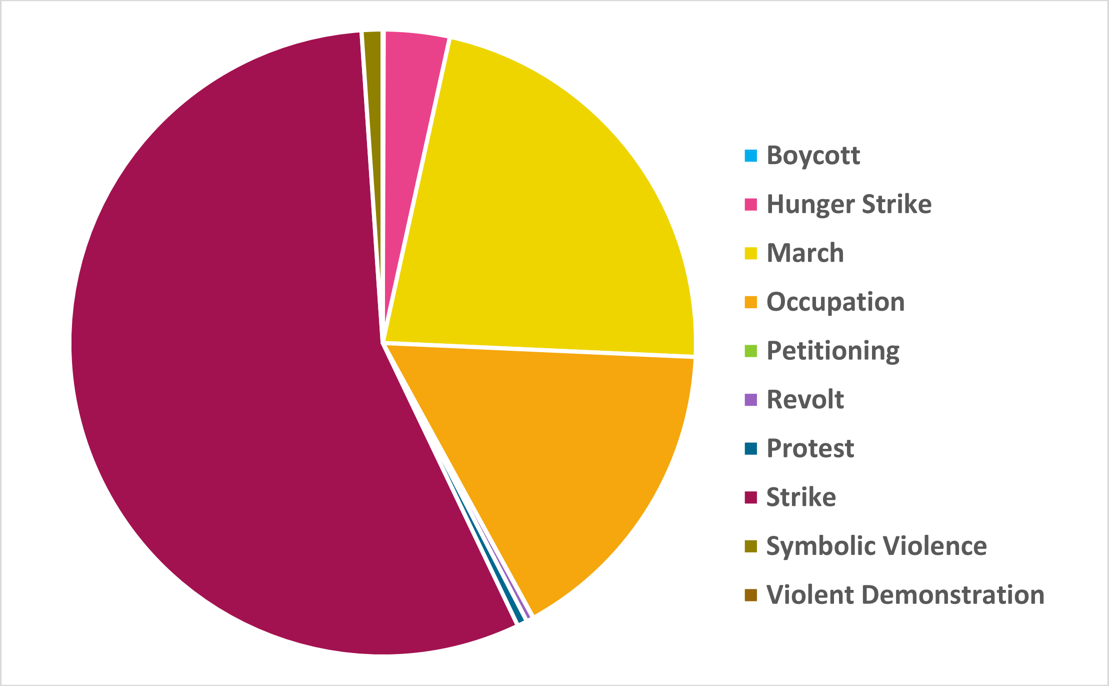


Figure 8: Protest Events Distribution

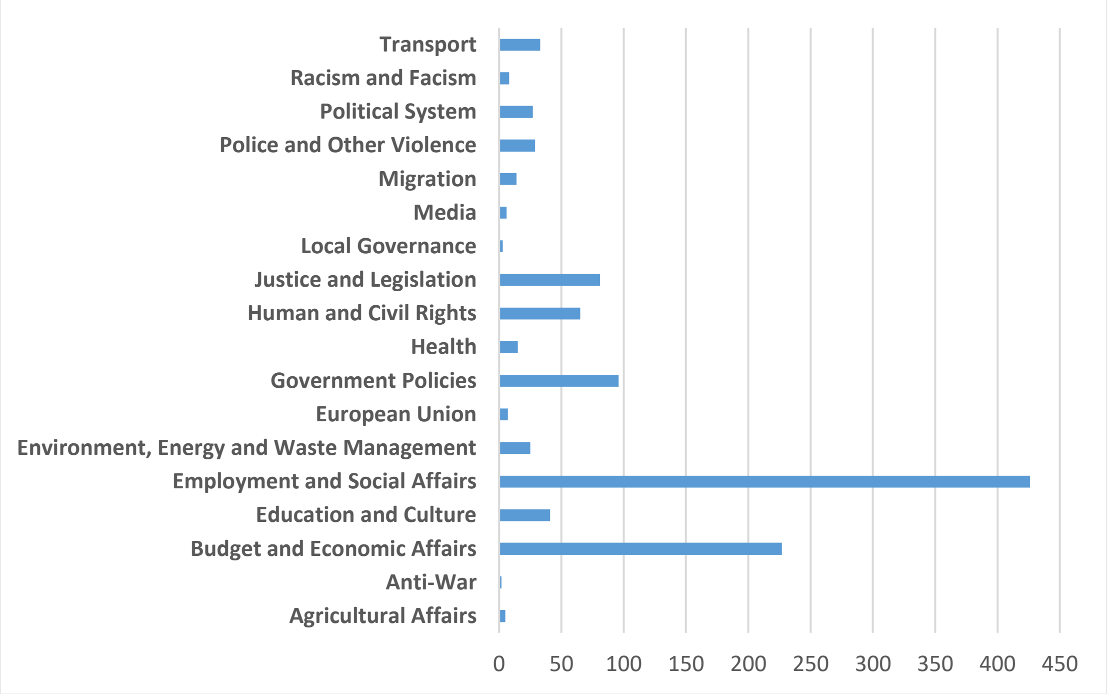


Figure 9: Issue Categories of Protest Events

6. Conclusions

The scope of this report was to present the Event Database that was produced in the context of the project "Examining xenophobia in Greece during the economic crisis: A computational perspective", along with the process followed for its generation. In accordance with the literature relevant to Event Extraction, an innovative methodology was implemented, with the most important element being the fact that it incorporated human-in-the-loop. Taking into consideration the fact that the work was interdisciplinary, involving both political scientists and computational experts, the exchange of knowledge was an integral part of the methodology. This was naturally an interactive process and resulted in a Codebook describing in details the expected outcome of the analysis. Several tools and technologies were then built and used for the computational implementation of the Codebook. The automatic analysis of the bulk of data collected, led to the population of a large Event Database, which was also described above and is available for further analyses through the CLARIN infrastructure.

Of course, several issues arose during the process of generating the Event Database. The first and maybe obvious difficulty concerned building a common ground between people coming from different disciplines. This challenge was overcome by close and frequent interaction.

Moreover, several limitations related to Natural Language Processing and resulting in errors recorded in the Database emerged. These inaccuracies appertain to three major categories. First, issues related to raw data wrangling, such as misspellings, typos as well as Optical Character Recognition (OCR) application errors during the automated conversion of raw input into machine readable text. Then, some pre-processing errors were detected, mainly related to the morphologically rich and syntactically complex nature of the Greek language. Finally, every system which automatically processes human language faces challenges associated with language complexity, like semantic ambiguity, one of the inherent characteristics of language.

As an extension of the above presented work, the enrichment of the Event Database using more event categories, constitutes the future aspirations of the project team. Moreover, it is the Natural Language Processing Lab’s constant ambition to evolve and enhance the developed systems so as to produce the best results.

7. References

Allan, J., J. Carbonell, G. Doddington, J. Yamron & Y. Yang. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription Workshop*.

Bond, D., Bond, J., Oh, C., Jenkins, J., Taylor, C. (2003). Integrated Data for Events Analysis (IDEA): An Event Typology for automated events data development. *Journal of Peace Research,* 40(6), 733-745.

Boschee, E., Natarajan, P., Weischedel, R. (2013). Automatic Extraction of Events from Open Source Text for Predictive Forecasting. In Subrahmanian V. (ed) *Handbook on computational approaches to Counterterrorism*. New York: Springer.

Boutsis, S., Prokopidis, P., Giouli, V., Piperidis, S. (2000). A Robust Parser for Unrestricted Greek Text. In *Proceedings of the 2nd Language Resources and Evaluation Conference*. Athens, Greece.

Chinchor, N., Marsh, E. (1998). MUC-7 Information Extraction Task Definition. In *Proceedings of the Seventh Message Understanding Conference*, MUC-7.

Filatova, E. & Hatzivassiloglou, V. (2003). Domain-Independent Detection, Extraction and Labeling of Atomic Events. In *Proceedings of the Fourth International Conference on Recent Advances in Natural Language Processing* (RANLP-2003). Borovets, Bulgaria.

Filatova, E. & Hovy, E. (2001). Assigning time-stamps to event clauses. In *Proceedings of the workshop on Temporal and Spatial Information Processing*, ACL. Toulouse, France.

Gerner, D., Schrodt, P., Yilmaz, O., Abu-Jabr, R. (2002). *Conflict and Mediation Event Observations (CAMEO): a new event data framework for the analysis of foreign policy interactions*. In Annual Meeting of the International Studies Association.

Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F. (2011). An Overview of Event Extraction from Text. In M. van Erp, W. R. van Hage, L. Hollink, A. Jameson, and R. Troncy, (eds), *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web at Tenth International Semantic Web Conference*. 779, 48-57.

King, G., Lowe, W. (2003). *An automated information extraction tool for international conflict data with performance as good as human coders: a rare events evaluation design*. International Organization 57(3), 617-642.

Leetaru, K., Shrodt, P. (2013). *GDELT: Global Data on Events, Language and Tone, 1979-2012*.

Li, Z., Wang, B., Li, M., Ma, W. (2005). A Probabilistic Model for Retrospective News Event Detection. In *SIGIR* 05.

O' Brien, S. (2012). A multi-method approach for near real time conflict and crisis early warning. In Subrahmanian V. (ed) *Handbook on computational approaches to Counterterrorism*.

Papageorgiou, H., Papanikolaou, K. (2017) Data Analytics meets Social Sciences: the Promap project. In Stathopoulou T. (ed.) *Transformations of protest in Greece.* (provisional title). Papazisis publishers, Athens.

Papageorgiou, H., Prokopidis, P., Demiros, I., Giouli, V., Konstantinidis, A. and Piperidis, S. (2002). Multi-level XML-based Corpus Annotation. In *Proceedings of the 3rd Language Resources and Evaluation Conference*. Las Palmas, Spain.

Papageorgiou, H., Prokopidis, P., Giouli, V. and Piperidis, S. (2000). A Unified POS Tagging Architecture and its Application to Greek. In *Proceedings of the 2nd Language Resources and Evaluation Conference*. 1455–1462. Athens, Greece. European Language Resources Association.

Papanikolaou, K., Papageorgiou, H., Papasarantopoulos, N., Stathopoulou, T. and Papastefanatos, G. (2016). “Just the Facts” with PALOMAR: Detecting Protest

Events in Media Outlets and Twitter. In *The Workshops of the Tenth International AAAI Conference on Web and Social Media*.

Prokopidis, P., Desypri, E., Koutsombogera, M., Papageorgiou, H. and Piperidis, S. (2005). Theoretical and practical issues in the construction of a Greek Dependency Treebank. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*. Barcelona, Spain.

Prokopidis, P., Georgantopoulos, B. & Papageorgiou, H. (2011). A suite of NLP tools for Greek. In *The 10th International Conference of Greek Linguistics*. Komotini, Greece.

Prokopidis, P., Papageorgiou, H. (2014). Experiments for Dependency Parsing of Greek. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*. Dublin, Ireland.

Pustejovsky, J. (2000). Events and the semantics of opposition. In C. Tenny and J. Pustejovsky (eds), *Events as grammatical objects*. Stanford, CA: CSLI Publications, 445–482.

Schrodt, P., Yilmaz, O., Gerner, D., Hermreck, D. (2008). *The CAMEO (Conflict and Mediation Event Observations) actor coding framework*. Paper presented at the International Studies Association, San Francisco.

Shrodt, P., Shannon, D., Weddle, J. (1994). Political Science: KEDS-A Program for the Machine Coding of Event Data. In *Social Science Computer Review*.

Wang, W., Zhao, D. (2012). Ontology-Based Event Modeling for Semantic Understanding of Chinese News Story. *Natural Language Processing and Chinese Computing*, 58-68. Springer.

Wueest, B., Rothenhäusler, K. and Hutter, S. (2013). Using computational linguistics to enhance protest event analysis. Annual Conference of the Swiss Political Science Association. Zurich: University of Zurich.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald B. T. & Liu, X. (1999). Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval*, 14(4), 32-43.