Proceedings of AESPEN 2020, pages 57–62
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Protest Event Analysis: A Longitudinal Analysis for Greece

Konstantina Papanikolaou[1,2], Haris Papageorgiou[1]
[1] Institute for Language and Speech Processing/Athena RC, Athens, Greece
[2] Omilia – Conversational Intelligence, Athens, Greece
kpapanikolaou@omilia.com, haris@athenarc.gr

Abstract

The advent of Big Data has shifted social science research towards computational methods. The volume of data that is nowadays available has brought a radical change in traditional approaches due to the cost and effort needed for processing. Thus, interdisciplinary approaches are necessary to cope with knowledge extraction from heterogeneous and diverse data sources. This paper presents our work in the context of protest analysis, which falls into the scope of Computational Social Science. More specifically, the contribution of this work is to describe a Computational Social Science methodology for Event Analysis. The presented methodology is generic in the sense that it can be expanded and applied in every event typology and moreover, it is innovative and suitable for interdisciplinary tasks as it incorporates the human-in-the-loop. Additionally, a case study is presented concerning Protest Analysis in Greece over the last two decades. The conceptual foundation lies mainly upon claims analysis, and newspaper data were used in order to map, document and discuss protests in Greece in a longitudinal perspective.

Keywords: Protest Event Analysis, Event Extraction

1. Introduction

Event Extraction has been a challenging task both for the field of Information Extraction in NLP and for Political and Social Sciences. As far as the latter is concerned, there have been several attempts to document events from news outlets, most of which were manual or semi-automatic.

The aim of this paper is to present an innovative computational methodology for the extraction of Protest Events from news data. Protest Event Analysis (PEA) has long been considered a significant tool for political scientists in the study of social movements and contentious politics (Wueest et al., 2013). Moving from tedious and time-consuming manual approaches used in this context, we implemented an automated methodology leveraging Natural Language Processing tools. We describe a Computational Social Science methodological approach to the research of PEA. More specifically, having Greece as reference, a longitudinal analysis of protests as a social phenomenon is documented and the impact of major socio-political events, like the recent economic crisis, is examined. Greece has been plagued by a severe financial crisis since the late 2009.

The work presented hereafter goes beyond traditional empirical approaches of social science research, thus aiming at analysing protest events using computational methods and big data analytics, exploiting a vast amount of available textual data from media outlets. We build upon an ecosystem of advanced computational content analytics technologies, capable of analysing large amounts of documents. Such topics, like PEA, are traditionally approached via small-scale, costly and non-reproducible expert coding of available political documents. However, the requirement of expert judgements is prohibitive in terms of cost and also restrictive in terms of the number of documents that could be analysed. Instead, the adopted methodology essentially develops an event database linking the major actors involved.

Therefore, a data analytics workflow was used to produce the corresponding data insights that allowed for the analysis of the complex issue of PEA and its evolution. Event analysis was performed, using news data from 2 different sources spanning the last two decades. The goal was to capture events correlated to protests along with the involved actors and record them into a large event database.

The paper is structured as follows: Related Work is discussed in Section 2. Event Extraction methodology is described in Section 3 and the Event Database in Section 4. The Evaluation of the developed system is presented in Section 5, while an Error Analysis is recorded in Section 6. Finally, the Results along with some valuable remarks are delineated in Section 7.

2. Related Work

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 et al., 1994). Evaluations have shown that hand coded and automatic coding of events 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 source for events. 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 et al., 2013). Despite its importance, the field of social protest in Greece is an almost uncharted territory and the related works are rather few (e.g. Kousis, 1999). Moreover, these studies are limited in their scope since they either cover a short timespan or are restricted to a specific topic (i.e. environment). This is partly due to the time-consuming nature of Protest Event Analysis (PEA), since, with a few exceptions (e.g. Imig and Tarrow, 2001, Wueest et al. 2013, Francisco n.d.), the identification and coding of protest events is done manually. The most important constraint of PEA method is the time needed for coding as the researchers have to read through literally

thousands of newspaper articles and then manually record all instances of protest events. Thus, most of the projects mentioned make use of a considerable amount of resources in terms of human capital and time.

3. Event Extraction Methodology

The framework that was designed and implemented for the Event detection task, is data driven and comprises five distinct steps, namely: (a) **Events Coding**: design of a taxonomy covering a wide spectrum of protest events, (b) **Data Collection**: a significant dataset was built from several news sources, (c) **Data Exploration** where humans were involved to provide valuable insights and create targeted data collections, (d) **Data Analysis,** the main phase of the task, during which the event database was populated, (e) **Data Visualization**, 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. Each of the aforementioned stages is further illustrated below.

3.1 Events Coding

The first step for the Protest Events Extraction task was the knowledge representation, namely the design of a coding schema encompassing a taxonomy of protest events. This task was undertaken by social and political scientists who, in collaboration with computational scientists, developed a Codebook (Papanikolaou et al., 2016) that incorporated several event types within the broader sense of protest events along, like *Strike*, *Hunger Strike*, *Demonstration*, *Blockade* e.tc. The Codebook was based on the Political Claim Analysis (PCA) research (Stathopoulou et al., 2018), thus the analysis unit is a Claim made in the public sphere, which comprises six distinct elements: *Form*, *Actor*, *Addressee*, *Issue*, *Location*, *Time*. In Information Extraction terminology, a Claim is an Event tuple consisting of six information types, i.e.:

1. *Form* is an event type depicting a way of action, like Boycott. This is an integral part of every event instance and all the other elements are connected to it.
2. *Actor* is the entity (person or organization) that acts, performs the action.
3. *Addressee* is the entity (person or organization) that is the target of the action, to whom the action is addressed.
4. *Issue* denotes the subject matter of a protest event, namely what the protest is about.
5. *Location* is the place where a protest event took place, and,
6. *Time* depicts the time the event happened.

In order for an event to be recorded in the Event Database, the necessary elements were *Form* and one of {*Actor*, *Addressee*, *Issue*}. Moreover, the entities denoting the *Actor* or the *Addressee*, were further classified into categories representing their role or status, for example *Government*, *Asylum seekers*, *Police*, *Tertiary Trade Unions* etc. Finally, the Issue information type was categorized in pre-defined topic classes, such as *Human and Civil Rights*, *Taxation and Fiscal Policies*, *Education* etc.

Therefore, each record in the Event Database comprises of the six aforementioned constituents and their attributes. Nevertheless, it is quite common that not all of the tuple elements are completed, according to the limitations mentioned above.

3.2 Data Collection

For the Event Extraction task, a large collection of news data was used. Specifically, the dataset comprised articles published in two nationwide newspapers with different political orientation, i.e. Kathimerini, a right-oriented and Avgi, a left-oriented paper; particularly, the articles included in the Wednesday and Friday issues were collected, for the time period spanning 1996-2014. All the articles are in Greek and also metadata-like section labels, headlines and the names of the authors were gathered along with the text itself. Hence, in total 540.989 articles, 314.527 from Kathimerini and 226.462 from Avgi were collected, prepared and stored. Data preparation included tackling 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, it sets out to incorporate human-in-the-loop. Therefore, human experts explored the collected dataset using queries. The aim of this process 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 in more complex ones with the use of Boolean operators. The exploration stage was also crucial for filtering the collected bulk of data and grouping them into event-oriented data clusters. This process was interactive and followed several iterations, as it was directed by the Codebook, which was also adjusted and enriched in line with the results of exploration.

One of the main goals of the Explorative Analysis was to better understand and obtain a wide view of the whole dataset. Given that the dataset consisted of two media sources reflecting ideological and idiosyncratic characteristics, it was essential to examine the different ways and linguistic means used by each news agency to report the same event. To this end, a full text search application for automated and scalable data processing was developed and used to index 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 as a new dataset to be further processed. They are also able to come back to the queries and modify them. Subsequently, in the data analysis phase, the saved queries along with the articles indicated

as relevant were retrieved and stored in data clusters, one for each event type.

3.4 Data Analysis

Event Extraction is a multifaceted task (Stathopoulou et al. 2018) since several information types are involved, which need to be detected in the text and interlinked. Overall, the adopted framework was data-driven and linguistically oriented. Its foundations lay on political and social sciences, additionally incorporating human-in-the-loop. The followed workflow first detects the structural components of the event and then links them to populate the event tuples which are then recorded in the Event Database. The employed methodology is semi-supervised, in the sense that a small fraction of data was labelled and used for the system development. Additionally, it is linguistically driven, thus morphosyntactic information from basic NLP tools is utilized to identify the information types defined in the Codebook.

The general workflow for extracting events is a pipeline in the sense that every module builds over the annotations produced by previous modules (Papageorgiou and Papanikolaou, 2017). At the first step, the ILSP-NLP tools suite (Papageorgiou et al., 2002; Prokopidis et al., 2011) was leveraged for pre-processing raw text and producing annotations for Tokens, Lemmas, Chunks, Syntactic relations and Named Entities. The next module of the pipeline is the Event Analysis Unit (EAU), which takes as input the output of the pre-processing phase and at first it detects the structural elements of the event and then uses linguistic rules based on shallow syntactic patterns to link the components and create an event tuple, recording and storing it in the Event Database (Pontiki et al. 2018). The Event Extraction system is a Finite State Transducers (FSTs) cascade, implemented using Gate JAPE patterns (Cunningham et al., 2000). Figure 1 depicts the Data Analytics stack for Event Extraction:

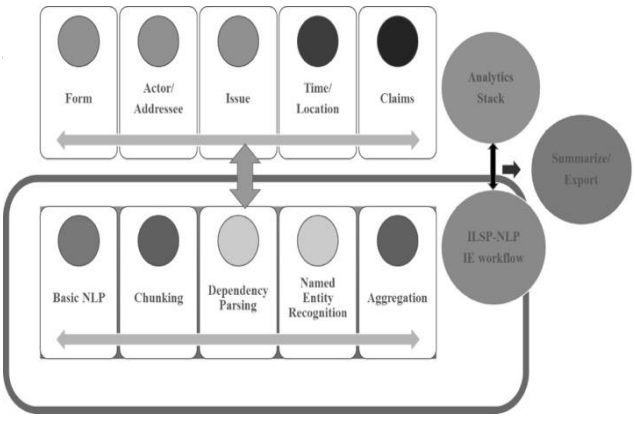

Figure 1: Data Analytics Stack

The above presented NLP workflow is fed with textual data. The basic (NLP) workflow includes segmentation (i.e. recognition of paragraph, sentence and token boundaries), part of speech tagging (i.e. assigning morphosyntactic categories to individual tokens), lemmatization (i.e. determining the base form of a token; both strike and strikes are attributed to the lemma strike), chunking (i.e. performing a shallow syntactic parsing and discovering syntactic constituents such as nominal and prepositional Phrases), parsing (i.e. determining the syntactic structure of each sentence) and Named Entity Recognition and Classification (NERC) identifying and classifying named entities into four major categories: Person, Organization, Location and Facility. This output is then forwarded to the EAU whose workflow is based on linguistic rules, given that semantics and shallow syntactic parsing patterns are exploited. EAU comprises several modules which seek to detect the structural components of the claim and to build links among them. Thus, first nominal lexicalizations of entities are identified and assigned the label *Candidate* along with Person and Organization annotations. After that, Time and Issue annotations are detected, while another module handles the identification of Forms. It is important to note that the Issue, namely the subject matter of the protest, is heavily depending on semantics. Consequently, patterns containing trigger words along with their syntactic complements were used for its detection. In such a pattern, a trigger word is "protest" and its syntactic complement a prepositional phrase starting with "about". Next, the pipeline decides whether an entity (named or nominal reference) can be assigned the label Actor or Addressee. At the final stage, the above annotations are extracted into the Event Database. The presented workflow is illustrated by the following indicative example. Given the following sentence:

The Law Society of Piraeus decided to occupy the Mortgage Registries of Piraeus and Salamis, on April 26th and 27th 2006, in protest against the serious operational problems it faces

the extracted output tuple recorded in the Database would be:

<**Actor**: *Law Society of Piraeus*, **Form**: *decided to occupy*, **Addressee**: *Mortgage Registries of Piraeus and Salamis*, **Issue**: *serious operational problems it faces*, **Time**: *April 26th and 27th 2006*, **Location**: *Piraeus, Salamis*>.

3.5 Data Visualisation

The Visualization phase is an integral part of the task as the results need to be visualized in different ways, making them understandable and easily perceivable 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 illustrative visualizations produced in the context of this work, are presented in the next section.

4. Event Database

The above presented methodology resulted in the population of the Event Database. More specifically, two files were created, one for each newspaper under examination, and then all the results were aggregated into one single database incorporating all the extracted event instances from both data sources. The database comprises several tables including the main information types and their attributes as were presented above. Moreover, there are tables recording metadata information. All the tables are linked using a unique ID as key.

5. Evaluation

The evaluation of our system was performed in two different ways. At first, a fraction of data was used, specifically the results of the *Strike* event type – which was the most prominent – and a time span of a month, 2/2014. The data were manually annotated, and the results compared to the system's output. The evaluation metrics used were *Precision* and *Recall*. For the selected data, Precision was 90% and Recall 93%.

Moreover, we conducted an extrinsic evaluation using data from GDELT, using event type Strike and Boycott which was part of the event coding used in our work. Since, data sources were different, the comparison was made on the basis of the recorded events in the timeline that coincided for both databases. The results can be seen in the following diagram.

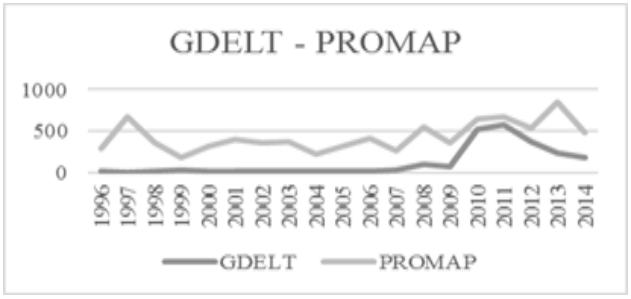

Figure 2: GDELT vs PROMAP results

6. Error Analysis

As mentioned above, the evaluation of the developed Event Analysis system showed significant results both in terms of precision and recall. Regarding recall, more experiments are needed for a more extensive evaluation, however taking into consideration the volume of the analysed data this is a quite tedious task. Despite this difficulty, at a small-scale evaluation, our system achieved a recall higher than 90% and at a large scale showed that the coverage of the events under examination is much better than GDELT, which is of great importance considering that there are no other similar analyses for Greek data. 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 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.

7. Results - Remarks

Both quantitative and qualitative observations emerge from the analysis of the results recorded in the Protest Event Database. In an initial statistical analysis examining the total number of Claims recorded in the Event Database, we made two remarks. First, the lowest number of protest events was documented in 2004 (Fig. 3), a year of relevant economic and social prosperity when Greece drew quite a lot of attention due to the Olympic Games held in Athens, which constituted a source of national pride. Additionally, it is clear that the total number of protest events indicates an increase after 2009, when the economic crisis first ensued.

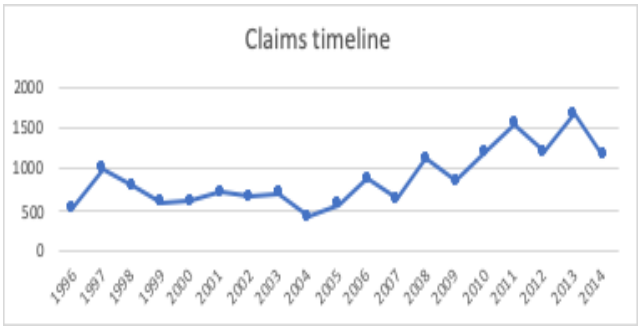

Figure 3: Total number of Claims

The top three event types in terms of frequency, were proven to be Strikes, Demonstrations and Occupations, indicating the ways the Greeks choose to protest and express their discontent (Fig. 4).

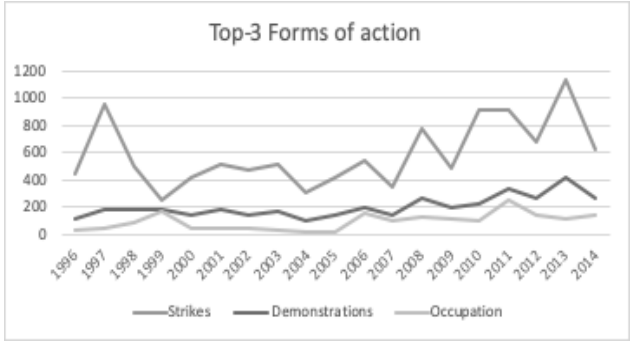

Figure 4: The top-3 forms of action

Finally, considering the most frequent topics under which the issues of the protests -taking place in the country for the examined time period- fall, it is obvious that the major concerns of the people are related to their economic status and employment affairs (Fig. 5).

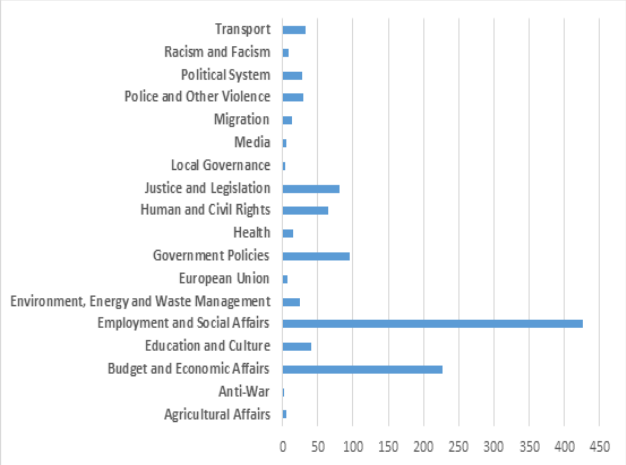

Figure 5: Issue Topic Categories

In addition, a qualitative analysis allows for some interesting observations. One of the most notable ones is the correlation between the number of recorded protest events and the election years. More specifically, looking at the chart in figure 3, we notice that the low spikes occur in election years. In particular, 1996, 2000, 2004, 2007, 2009, 2012 were all years of national elections and it is clear that the total number of protests during those years, show a significant decrease. Nevertheless, as computational scientists we can only point out a correlation, but it is designated to political scientists to interpret such phenomena (Stathopoulou et al., 2018).

8. Conclusions

In this paper, an automated approach for Protest Event Extraction was presented. In accordance with the literature relevant to Event Extraction, an innovative methodology was implemented, with one of the most prominent elements 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. The development processes along with the database were described above in detail.

As an extension of the above presented work, the enrichment of the Event Database using more socio-political event categories, constitutes the future aspirations of the team. Moreover, it is our constant ambition to evolve and enhance the developed systems so as to produce the best results.

9. Acknowledgements

We acknowledge support of this work by the project "Computational Science and Technologies: Data, Content and Interaction" (MIS 5002437) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

10. Bibliographical References

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.

Cunningham H., Maynard D. and Tablan V. (2000). JAPE: a Java annotation patterns engine. *Research Memorandum CS–00–10*, Department of Computer Science, University of Sheffield.

Francisco, R. n.d. *European Protest and Coercion Data*. Available from: http://web.ku.edu/~ronfrand/data/. [retrieved 12 February 2016].

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.

Imig, D. and Tarrow, S. (eds). 2001. *Contentious Europeans Protest and Politics in an Integrating Europe*. Lanham: Rowman & Littlefield.

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.

Kousis, M. 1999. Environmental Protest Cases: The City, The Countryside, and The Grassroots in Southern Europe. In *Mobilization* 4(2): 223-238.

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

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. and Papanikolaou, K. (2017). Data Analytics meets Social Sciences: the Promap project. In Stathopoulou T.(ed.) *Transformations of protest in Greece*. 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.

Papanikolaou, K., Papageorgiou, H., Papasarantopoulos, N., Stathopoulou, T., Papastefanatos, G. (2016). *"Just the Facts" with PALOMAR: Detecting Protest Events in Media Outlets and Twitter*. In International AAAI Conference on Web and Social Media. North America.

Prokopidis, P., Georgantopoulos, B. and Papageorgiou, H. (2011). A suite of NLP tools for Greek. *In Proceedings of the 10th International Conference of Greek Linguistics*, Komotini, Greece, pp. 373–383.

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

Stathopoulou, T., Papageorgiou, H., Papanikolaou, K., Kolovou, A. (2018). Exploring the dynamics of protest with automated computational tools. A Greek case study. In *Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications*. German Society for Online Research.

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.