Comparative Agendas Project: Polish dataset

Key words: public policy, comparative research, data coding, datasets (social science)

The main objective of the project is to set up the Polish branch of the Comparative Agendas Project (CAP), accordingly with its coding scheme.

CAP provides researchers, students, decision makers and other interested users with a free access to data relevant for tracking public policy processes across time and space. Currently, CAP covers the following countries: Australia, Belgium, Brazil, China, Croatia, Denmark, France, the Netherlands, Spain, Israel, Canada, Germany, New Zealand, Portugal, Russia, Hungary, Italy, Switzerland, Turkey, USA and the United Kingdom. Some data for the state of Florida and Pennsylvania and the European Union are also available. Poland, still, is a blind spot here.

CAP classifies political activity (e.g. parliamentary debates, leaders' speeches, committees hearings, legislative initiatives, laws enacted, judicial decisions) in a unified and consistent coding system. It consists of 21 major categories and over 220 minor topics. The main tier of the scheme includes the following policy areas:

Major Topic	Title
1	Domestic Macroeconomic Issues
2	Civil Rights, Minority Issues, and Civil Liberties
3	Health
4	Agriculture
5	Labor and Employment
6	Education
7	Environment
8	Energy
9	Immigration and Refugee Issues
10	Transportation
12	Law, Crime, and Family Issues
13	Social Welfare
14	Community Development and Housing Issues
15	Banking, Finance, and Domestic Commerce
16	Defense
17	Space, Science, Technology, and Communications
18	Foreign Trade

19	International Affairs and Foreign Aid
20	Government Operations
21	Public Lands, Water Management, and Territorial Issue
23	Cultural Policy Issues

All the CAP data are available in open access formula at https://www.comparativeagendas.net/datasets_codebooks.

CAP dataset already enables studying public policy agenda in at least two research designs: time series analysis and comparative approach. Thus, some further developments are envisaged through merging the CAP data with other sets, e.g. election results, roll calls, political party platforms (Manifesto Research Project), government occupation (ParlGov Database), public opinion polls or economic indicators. All in all, this allows for gathering a substantial policy-relevant empirical data.

Comprehensive and multidimensional CAP dataset is also worth considering for **machine learning applications in political science**. Notwithstanding developments in computer science, this is still relatively undeveloped research area in political science. Thus, publication in highly cited venues (e.g. *Policy Studies Journal*, *Journal of European Public Policy*) is a viable option as witnessed by the forthcoming publication on the supervised learning-based classification by Miklos Sebok and Zoltan Kacsuk in the top methodology journal of the field (Political Analysis, Impact factor: 4.2) and the Palgrave volume on Hungarian CAP data (forthcoming). For the above substantive reasons and due to international project team, any results will be submitted to international (i.e. foreign, neither Hungary- nor Poland-based) journals.

As for details, please refer to the below list of selected variables available currently for the United States:

1. Parliamentary & Legislative:

Congressional Bills

Congressional Hearings

Congressional Research Service Reports

Public Law Titles

Public Laws

Roll Call Votes

2. Prime Minister & Executive:

Executive Orders

Presidential Veto Rhetoric

State of the Union Speeches

3. Judiciary:

Supreme Court Cases

4. Budget:

Budget Authority (Adjusted)

Budget Authority-Policy Crosswalk

Budget Outlays

Tax Expenditures

For obvious reasons, the above list is only illustrative for possibilities behind the CAP. It is acknowledged that making anything similar in scope to the USA dataset calls for a research project that spans well beyond 12 months envisaged in the current call.

The main objective of the project is to classify data already gathered by the Jagiellonian Center for Quantitative Research in Political Science (Centrum Badań Ilościowych nad Polityką) on the Polish parliament and government accordingly with the CAP coding scheme. The following data is currently available for coding:

- bills (legislative proposals) (6,000 +),
- Sejm agenda (16,000 +),
- parliamentary debates (330,000 +),
- committee hearings (10,000 +),
- interpellations, parliamentary questions, and ministerial statements (220,000 +),
- reports submitted to the parliament by the government (ca. 800),
- motions of no confidence (ca. 100),
- Council of Ministers legislative proceedings (6,000 +),
- Constitutional Tribunal complaints (800 +) and decisions (2,500 +),
- laws in force (ca. 750).

Each of the above examples of political activity may be coded by the CAP scheme. This, however, is not viable for hand-coding due to the volume of data points available. Thus, **machine learning classification** methods are envisaged such as topic identification and

segmentation in Natural Language Processing (NLP). Currently, it is envisaged that the following approaches will be harnessed:

- a convolutional neural network with a hand-coded sample of training examples,
- key words-based SVM classifier,
- community detection in data cross-reference graphs.

Non-textual variables (committee references, identification of ministry responsible, etc.) will be used to assign data points to groups with distinct Bayesian priors. Here, the input from the investigator who led the setup of the Hungarian CAP project (cap.tk.mta.hu) and more than a dozen individual datasets would be of a great importance.

Additionally, a random sample of data will be manually hand-coded by experts. This not only allows for training supervised learning algorithms but comparing human- and machine-coding approaches in terms of their performance and to identify areas in need of improvement.

Relevance for the DigiWorld Priority Research Area

The project's relevance for the DigiWorld Priority Research Area stems from the **application of machine learning approaches for coding qualitative data** in accordance with the CAP scheme. This implies application of machine learning methods in social science and digital humanities through the use of NLP and quantitative text analysis. Furthermore, this dataset could be used as a departure point for **further developments in applications of machine learning techniques in political science.**

The 4I Principle

Interdisciplinarity: application of methods developed in mathematics and computer science in political science seem to warrant the point. Also, the projected dataset would be available to researchers from other scientific disciplines. As of today, the CAP data was already used in e.g. economy, media research, sociology, social psychology and management.

Internationalisation: the project is aimed at building the Polish branch of international Comparative Agendas Project. Also, the team includes Dr. Miklós Sebők who co-founded the Hungarian branch of the CAP and has been leading its research projects ever since.

Innovation stems from the application of the proposed methodology, i.e. NLP text classification. This approach is still relatively sparsely represented in political science

academia. What is more, the use of network analysis of relations between laws, reports and debates to assist classification is a novel idea.

Integration: the obvious target is a research community. This, however, does not preclude using project's results by other communities such as *think tanks*, NGOs, decision makers, policy commentators, whistle-blowers and media reporters. Data gathered could be used for evidence-based policymaking and policy tracking – both areas are far from being irrelevant.

Expected results

The main objective is setting up the database that could be part of the CAP framework. Also, the project would result in an article describing methodology, obstacles and opportunities behind qualitative coding of random sub-samples. Our target venue would be one of the leading journals such as *Policy Studies Journal* or *Journal of European Public Policy* (both worth 140 points) or *Journal of Computational Social Science* (70 points, but the journal is new with emerging IF potential).

Dataset users would be asked to cite that article as a reference.