Exploring the Political Agenda of the Greek Parliament Plenary Sessions

Dimitris Gkoumas, Maria Pontiki, Konstantina Papanikolaou, Haris Papageorgiou
Institute for Language and Speech Processing, Athena Research and Innovation Center
Artemidos 6 & Epidavrou, 15125, Athens, Greece
{dgkoumas, mpontiki, konspap, xaris}@ilsp.gr

Abstract

In this paper we present primary content analysis results for the Greek Parliament data in the context of Natural Language Processing and Text Mining approaches. The raw minutes of the Greek Parliament plenary sessions of the last 26 years are processed and transformed into a structured and machine readable format, and then clustered based on the analysis of their content using topic modelling techniques. Inspired by and following the work of Greene and Gross (2017) for the European Parliament, we employ a two-layer methodology for applying topic modelling in a Non-negative Matrix Factorization framework to a timestamped corpus of political speeches in order to explore dynamic topics. The results are visualized in various ways (by topic, by time) providing at the same time information about the contribution of each Parliament Member, political party and region (constituency) to each topic, and by extent, the ability to explore how the political and policy agenda has been shaped and evolved in Greece over time.

Keywords: Greek Parliament Data, Dynamic Topic Modelling

1. Introduction

Parliament plenary sessions depict the peak of the legislative work done by policy makers (members of the Parliament, government officials) and thus, their content constitutes a valuable source for the exploration of the way in which political and policy agendas are shaped and evolve over time (e.g. how problems and issues are defined, constructed, and placed on the political and policy agenda), as well as of the way in which the individual Parliament Members and political groups react and act over time (e.g. voting behavior). The recent work of Greene and Gross (2017) that analyses the content of the European Parliament plenary sessions using a dynamic topic modelling approach, indicates that the detection of latent themes in a timestamped corpus of political (legislative) speeches can provide insights of the way in which the political agenda reacts to exogenous events (e.g. the emergence of the Eurocrisis) and evolves over time. Such analysis can also supply information about the Parliament members' reactions to different stimuli. Different types of topic modelling approaches have been used in the political science literature also tracing the political attention of individual politicians over time based on the themes they speak about (Quinn et al., 2010), or capturing the political priorities expressed in Congressional press releases (Grimmer, 2010).

In this context, we present a work focusing on the Greek Parliamentary (GrParl) data. Inspired by and following the work of Greene and Gross (2017), we adopt their two-layer strategy for applying topic modelling in a Non-negative Matrix Factorization (NMF) framework to explore dynamic topics in the GrParl plenary sessions. The contribution of our work is two-fold: 1) Transformation (digitization/normalization/processing) of the GrParl plenary sessions minutes into a structured and machine readable format (Section 2) making them for the first time available for Natural Language Processing (NLP) and Text Mining approaches. 2) A platform that enables an insightful navigation across the GrParl plenaries based on the results of the topic modelling analysis (Section 3); the speeches can be explored by dynamic topics and by time. Information about the contribution of each GrParl member, political party and region (constituency) to each topic is also provided enabling a more perceptive monitoring of the political and policy agenda in Greece over time for all stakeholders (e.g. journalists, political & social scientists, policy makers).

2. Data Collection

The Greek Parliament provides the proceedings of all the plenary sessions during the time period from 1989 to 2015. The data was in different formats (e.g. txt/doc/docx/pdf/jpg). With the exception of the image format files covering the time period 1994-1996, the data was transformed into a structured and machine readable xml version, and organized based on its timestamp. In particular, the collection consists of 4356 plenary sessions containing in total 1.063.546 unique speeches.

2.1 Data Processing

The data was processed using our in-house Athena R.C./ILSP suite of NLP tools for the Greek language (Papageorgiou et al., 2002; Prokopidis et al., 2011). In particular, the text of each plenary session was enriched with the following types of information:

- Segment annotation: the start/end point of each speech.
- Speaker annotation: the name of each speaker and his/her political affiliation (the name of the party he/she represents in the Greek Parliament).
- POS tagging: the Part-of-Speech (e.g. adjective, noun) of each word in the text.
- Temporal annotation: the date of each speech.

In a next phase, the data was organized by the date of the speeches.

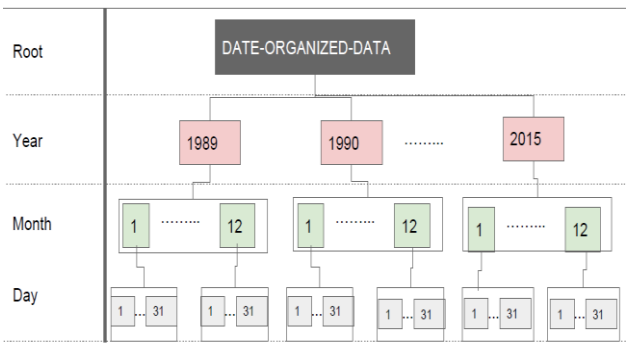

Figure 1: GrParl Data Organization.

2.2 Data Organization

Dividing temporal data into time windows of fixed duration when applying clustering techniques is often suggested in the literature (Sulo et al., 2010). After several experiments, we decided to set one month windows, and to have as many speeches as possible per window. In particular, all speeches made by Parliament members in the same month belong to the same window (see Figure 1). Thus, we ended up having 283 windows for the time period from 1989 to 2015. Hence, we were able to discover topics that were relevant to each month of each year as well as to monitor their dynamics over time.

3. Topic Modelling

Topic modelling is a widely used data analysis technique that provides an effective way to obtain insights in large collections of unlabeled data; topic models are used for inferring low-dimensional representations that uncover the latent semantic structures of different types of data; textual (Mcauliffe and Blei 2008), image (Wang, Blei, and Li 2009), or audio (Hoffman, Blei, and Cook 2009) data, among others. The state of the art topic modelling approaches are often based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non Negative Matrix Factorization (NMF) (Greene and Gross 2017; Kuang, Choo, and Park 2015). NMF algorithms provide semantically meaningful output that is easily interpretable in clustering applications (Kuang, Choo, and Park 2015).

In particular, the TF-IDF weighting scheme that the NMF is using enables the calculation of the importance of a word to each text in a collection of documents based on weighted term-frequency values, and thus, the generation of various but semantically coherent topics that are less likely to be represented by the same high-frequency terms. Such an attribute is important for models applied in the context of political speech in parliaments, as it provides the ability to "*differentiate between broad procedural topics relating to the day-to-day running of plenary and more focused discussions on specific policy issues*" (Greene and Gross 2017).

3.1 Dynamic NMF Method

Following the work of Greene and Cross (2017), we employed a Dynamic NMF approach for the GrParl data. In Dynamic NMF, the idea is that after identifying a certain number of topics in each 1-month-window, the ones which are semantically similar are grouped together as "dynamic topics", rather than identifying topics from the beginning. In a first phase, the NMF model runs on the time windows into which the data is divided and generates topics for each window. In a second phase, the NMF algorithm is applied again, with the number of topics decided after applying the Topic Coherence via Word2Vec (TC-W2V) measure proposed by O' Callaghan et al. (2015); TC-W2V evaluates the relatedness of a set of top terms describing a topic by computing a set of vector representations (word embeddings) for all of the terms in a large corpus using the Word2Vec tool introduced by Mikolov et al. (2013). For more details about the NMF method consult Greene and Cross (2017).

The input for the topic models was the preprocessed and organized data described in Section 2. In particular, each window corresponds to the monthly sessions and consists of the speeches of all Parliament members. During the construction of the document-term matrix the stop words were filtered out (using a relative list) and only the content words were kept (adjectives, nouns, adverbs and verbs). During the TF-IDF calculation rare terms that appeared in less than 10 documents were removed. Given that the TF-IDF scheme calculates the frequency of each term in each document, and also in the whole collection of the documents, the initially generated topics were somewhat noisy. In order to deal with this, units with length less than three lines were discarded and only the nouns and the adjectives were kept. In this way, we ended up with 283 comprehensive and coherent window-based topics and 90 dynamic topics. To detect the number of dynamic topics *k*, we used the TC-W2V coherence measure as in (Greene and Cross, 2017) using the GrParl speeches as a training corpus for the extraction of the word embeddings.

3.2 Presentation of the Results

The extracted topics are listed in a table, where for each dynamic topic a user can see its most descriptive terms, the number of parties and the regions (constituencies) associated with it, and its frequency (i.e. the number of the time windows in which it appears). For example, as illustrated[1] below in Figure 2, D87 is a migration related topic described by the terms: "immigrant" (μετανάστης), "nationality" (ιθαγένεια), "foreigner" (αλλοδαπός), "immigration" (μετανάστευση), "refugee" (πρόσφυγας), "asylum" (άσυλο), "illegal immigrant" (λαθρομετανάστης), "hellenism" (ελληνισμός), "society" (κοινωνία), "homeland" (πατρίδα), "reception" (υποδοχή), "language" (γλώσσα), "emigrant" (απόδημος), "integration" (ένταξη), "identity" (ταυτότητα), "criminality" (εγκληματικότητα), "victim" (θύμα), "government" (πολιτεία).

ID	Terms	#Parties	#Regions	Frequency
D87	μεταναστης, ιθαγενεια, μεταναστευτικος, αλλοδαπος, μεταναστευση, προσφυγας, ασυλο, λαθρομεταναστης, ελληνισμος, κοινωνια, πατριδα, υποδοχη, γλωσσα, αποδημος, ενταξη, ξενος, ταυτοτητα, εγκληματικοτητα, θυμα, πολιτεια,	18	58	61

Figure 2: Snapshot of the table presenting the dynamic topics.

[1] The results are available at the following link: http://194.177.192.82/presentation/index.html

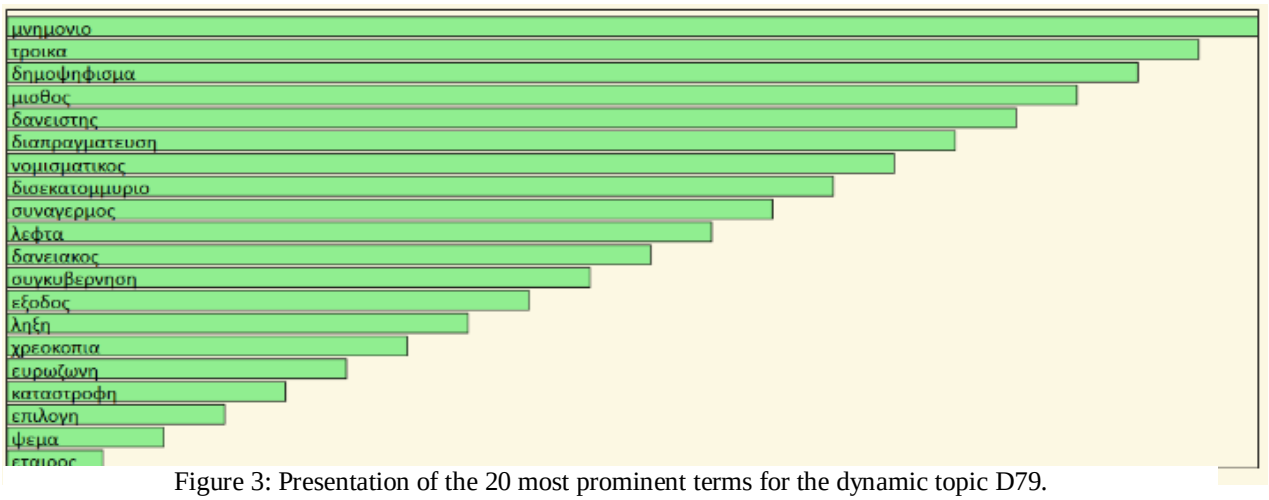

Figure 3: Presentation of the 20 most prominent terms for the dynamic topic D79.

By clicking on a row, users are able to study in detail the corresponding dynamic topic. In particular, the following types of information are provided:

- The 20 most prominent terms of each dynamic-topic (Figure 3). The words are presented in ascending order according to their importance, with the first one being the most descriptive term for each dynamic topic. For example, as illustrated above in Figure 3, the top term for the topic D79 is the word "memorandum" (μνημόνιο) and it is highly associated with the words "Troika" (Τρόικα), "referendum" (δημοψήφισμα), "salary" (μισθός), "creditor" (δανειστής), "negotiation" (διαπραγμάτευση), "monetary" (νομισματικός), etc.
- The most important monthly topics that are related to each dynamic topic. The window topics are listed in a table similar to the one in Figure 2, and presented in ascending order, with the first being the most important one. For each window topic, a user can see its timestamp, the most descriptive terms, the number of the politicians, parties and regions (constituencies) associated with it, as well as the number of the corresponding segments (speeches). By clicking on a table's row more information is available about the corresponding window topic.
- Information about the speakers that are related to each dynamic topic. In detail, users can see the speakers' names, the party they belong to, the region they are elected to represent and the date of each speech. By clicking on a table's row users can read or download the corresponding speech.
- A chart demostrating the contribution of each party to each dynamic topic (Figure 4).
- A chart demostrating the contribution of each region to each dynamic topic.

Users have also the option to explore the whole corpus of the GrParl plenary sessions based on the time windows that it has been divided to (283 in total) for the purposes of our analysis. As illustrated in Figure 5, for each time window we present the number of the topics, the terms, the parties, the politicians and the segments it is associated to. Again, by clicking on a table 's row, users are able to explore the corresponding window in more detail. Finally, users can use specific keywords as search terms to discover related topics.

4. Discussion

In this paper we present primary topic modelling results for the GrParl data. Following the two-layer NMF

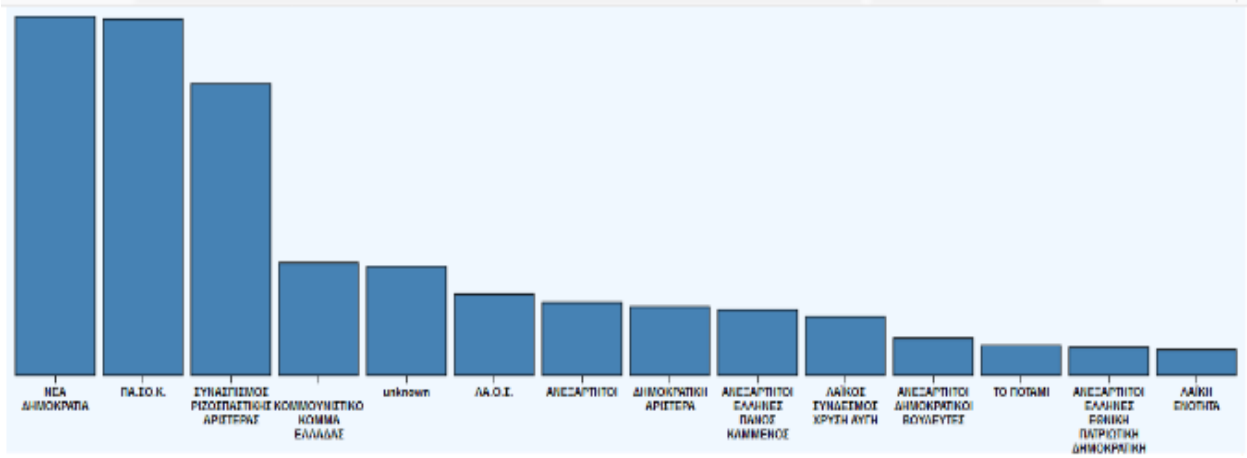

Figure 4: Chart demostrating the contribution of each political party to the dynamic topic D79.

Show 50 entries Search:

Time Window	#Topics	#Terms	#Parties	#Politicians	#Segments
2008_12	19	304	5	291	1216
2009_12	10	189	8	286	1369
2010_12	10	178	8	285	1187
2008_03	14	229	5	276	1654
2013_12	17	261	9	276	1444
2010_05	14	239	8	273	1523
2015_12	25	371	8	273	1327

Figure 5: Time-window based presentation of the GrParl corpus.

methodology proposed by Greene and Gross (2017) for identifying dynamic topics in large political speech corpora over time, we analysed the minutes of the GrParl plenary sessions during the time period from 1989 to 2015. More than one million speeches made by Parliament members during the last 26 years in Greece have been processed and analysed, and can be explored by topic and by time. Information about the contribution of each GrParl member, political party and region to each topic is also provided enabling a more insightful navigation across the GrParl plenaries. The next step of our work is the qualitative analysis of the extracted dynamic topics and the assignment of a descriptive label to each topic based on the most descriptive words appearing in speeches related to it.

In the context of an interdisciplinary work with political scientists, we plan to explore the evolution of selected dynamic topics of interest over time focusing on specific case studies, to examine their connection to significant events (e.g. refugee crisis, Eurocrisis), and to compare them with the corresponding ones in the European Parliament and, if possible, in other European countries. Future work also involves applying other types of content analysis techniques, for example Quotations Extraction and Sentiment Analysis, in order to enrich the original data with more insights for anyone interested to scrutinize the policy agenda as well as the policy makers' reactions and actions in Greece over time (e.g. journalists, political scientists, policy makers, stakeholders).

5. Acknowledgements

We acknowledge support of this work by the projects "Computational Science and Technologies: Data, Content and Interaction" (MIS 5002437) and "APOLLONIS, Greek Infrastructure for Digital Arts and Humanities, Language Research and Innovation" (MIS 5002738), which are 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). Part of the work reported here was made possible by using the CLARIN infrastructure**.**

6. Bibliographical References

Blei, D. M., Ng, A.Y., Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022.

Greene, D. and Cross, J.P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Political Analysis 18(1), 1–35.

Hoffman, M., Cook, P., and Blei, D. (2009). Bayesian spectral matching: Turning Young MC into MC Hammer via MCMC sampling. In: International Computer Music Conference.

Kuang, D., Choo, J. and Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Partitional Clustering Algorithms, pp. 215–243.

Mcauliffe, J.D., and Blei, D.M. (2008). Supervised topic models. In: Advances in neural information processing systems, pp 121–128.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. CoRR abs/1301.3781.

O'Callaghan, D., D. Greene, J. Carthy, and P. Cunningham (2015). An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications (ESWA) 42(13), 5645–5657.

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 International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain, pp. 1723–1728.

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.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. American J. Political Science 54(1), 209–228.

Sulo, R., T. Berger-Wolf, and R. Grossman (2010). Meaningful selection of temporal resolution for dynamic networks. In Proc. 8th Workshop on Mining and Learning with Graphs, pp. 127–136. ACM.

Wang, C., Blei., D., and Fei-Fei, L. (2009). Simultaneous image classification and annotation. In: Computer Vision and Pattern Recognition.