Investigating Political Ideologies through the Greek ParlaMint corpus

Maria Gavriilidou, Dimitris Gkoumas, Stelios Piperidis, Prokopis Prokopidis
ILSP / Athena RC
Artemidos 6, 15125 Marousi, Greece
{maria, dgkoumas, spip, prokopis}@athenarc.gr

Abstract

This paper has two objectives: to present (a) the creation of ParlaMint-GR, the Greek part of the ParlaMint corpora of debates in the parliaments of Europe, and (b) preliminary results on its comparison with a corpus of Greek party manifestos, aiming at the investigation of the ideologies of the Greek political parties and members of the Parliament. Additionally, a gender related comparison is explored. The creation of the ParlaMint-GR corpus is discussed, together with the solutions adopted for various challenges faced. The corpus of party manifestos, available through CLARIN:EL, serves for a comparative study with the corpus of speeches delivered by the members of the Greek Parliament, with the aim to identify the ideological positions of parties and politicians.

Keywords: parliamentary corpora, party manifestos, ideology identification

1. Introduction

Parliamentary data is considered extremely important as it contains rich linguistic content corresponding to local and international events, on political, social, economic, environmental and health issues, among others. In addition to the significance of the content, rich metadata (e.g., speaker, party affiliation, gender, role) as well as additional clues (interruptions, voting results) can often be obtained. In the field of political science, the study of political ideology and position (left-right) of members of parliament (MPs) is of great importance. For this reason, this paper aims to describe the process of assembling and encoding the ParlaMint-GR corpus as part of the ParlaMint project, and to investigate the ideology of politicians in the Greek parliament, based on two sets of corpora: the *ParlaMint-GR corpus*, on the one hand, and the *Party manifestos of Greek Parliamentary Parties corpus*, on the other. Both corpora are available through the CLARIN:EL infrastructure, while the ParlaMint-GR corpus is also available through the CLARIN.SI repository.

Section 2 describes the ParlaMint project, which provided the framework within which the ParlaMint-GR corpus was created; Section 3 elaborates on the creation of the ParlaMint-GR corpus, the solutions adopted for data acquisition, encoding and annotation; Section 4 describes the Party manifestos corpus; Section 5 presents the creation of ParlaMint-GR specific word embeddings; Section 6 discusses the experiments on the two corpora with the aim to investigate various aspects of political ideology declaration, and Section 6 concludes with future steps.

2. The ParlaMint Project

The objective of the ParlaMint project (Erjavec et al., 2022) was the creation of multilingual, comparable, and uniformly annotated corpora, following uniform collection principles, adhering to common structural and linguistic annotation principles and to a common metadata model. The first phase of the project (ParlaMint I: 2020 – 2021) produced 17 corpora, while the second phase (ParlaMint II: 2022 – 2023) resulted in corpora in 29 languages, one of them being Greek. All corpora were automatically translated into English for comparability purposes. They are hosted at the Slovenian CLARIN repository[1], accompanied by tools for querying the data (such as concordancers, corpus analysis and statistical tools, etc.) (Erjavec et al. 2023).

3. The Corpora

3.1 The ParlaMint-GR Corpus

3.1.1 Data Acquisition

The source for Greek parliamentary data acquisition was the Hellenic Parliament official site[2], the scraping of which yielded 1,263 files in total (approximately 50 MWs), which consist in approximately 350,000 speeches made by 634 members of the Parliament and corresponding to proceedings from January 2015 to February 2022[3]. Besides the speeches, the proceedings also contain transcribers' notes, which record various incidents happening during the Parliamentary meetings (e.g., notes related to time *"the meeting started at 10:00 am"* or recording voting results *"80 voted Yes and 57 voted No"* etc.), vocal non-lexicalized sounds such as shouts, laughter, etc., clapping, or any other incident affecting communication.

The Greek parliament is a unicameral parliament with a multi-party political system. The proceedings are organized in Parliamentary terms (a term is the period between two general elections). Each Parliamentary term is divided into Sessions; a parliamentary term has regular Sessions, while extraordinary and special Sessions are also foreseen. Each Session is divided

[1] https://www.clarin.si/repository/xmlui/handle/11356/1859
[2] https://www.hellenicparliament.gr/en/
[3] https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 116–120

into Meetings, and each Meeting into Sittings (multiple sittings are possible, e.g., morning/afternoon sittings).

3.1.2 Data Encoding and Metadata

The collected speeches made by members of Parliament and recorded in the proceedings were automatically processed. For every speech external metadata were provided, as well as structural and linguistic annotation. These tasks were preceded by a necessary phase of data and metadata curation, given that the minutes were not always free of typographical errors, spelling mistakes or discrepancies (e.g., in the names of members of parliament).

Dedicated metadata obtained from various sources were added for all relevant entities, i.e., government, political parties, and members of parliament (MPs). Metadata for each government, i.e. the starting and ending date of governance, Prime Minister and ministers of each government, their corresponding ministries with the relevant dates, as well as any resignations or suspensions, were obtained from the Secretariat General for Legal and Parliamentary Affairs[4], where this information is provided for all Greek governments since 1909.
For each party, the metadata include its name, its acronym, its leader, the year of establishment and the year it ceased to exist (where appropriate), whether it forms part of the government or the opposition, a link to its Wikipedia page (where additional information can be found), and finally the party's position as regards ideology and policy issues (based on Chapel Hill Expert Survey[5]). The Chapel Hill expert surveys estimate party positioning on European integration, ideology and policy issues for national parties in a variety of European countries. The first survey was conducted in 1999 (14 Western European countries), and the latest in 2019 (31 countries). The survey includes questions on various issues such as Ideology, EU integration, Specific EU Policy Questions (Agriculture, Environment, Economics, etc.), Tax policy, Welfare, immigration, position on civil rights, human rights etc., and places each party on a spectrum from far left to far right.

Information on the political party each MP belonged to during each period was acquired by scraping the Hellenic parliament official site, where a dedicated page[6] lists all Members of Parliament from 1974 until today, with their political affiliations. For additional information about each MP, such as their gender, their parliamentary roles (i.e., Prime Minister, Minister, party president, parliament president, vice-president, etc.), their government positions, and electoral districts (Dritsa, 2020)[7] was deployed, with modifications and adjustments to the original code.

For each of the 1,263 proceedings' files the following types of information were identified and annotated: term, number and date for each session, beginning and ending of each speech, speaker and his/her role (Chairperson, Regular speaker, Guest speaker). Each speech that was extracted from the Hellenic Parliament proceedings files was encoded as utterance. Each utterance was classified as being a proper Speech by an MP, or as Vocal, i.e., non-lexical vocal sounds by other MPs (shouts, laughter, etc.), as recorded in the minutes, and annotated accordingly.

The association of each speaker with the collected metadata (role, gender, political party and period) was based on Jaro-Winkler distance calculation between the speaker in question and all possible MPs in our list. The similarity threshold was set at 0.95, to avoid false positives.

3.1.3 Linguistic Annotation

Besides structural annotation and the relevant metadata, all proceedings files were automatically processed and linguistically annotated. For the linguistic processing we used the ILSP Neural NLP toolkit (Prokopidis and Piperidis, 2020), available through CLARIN:EL[8]. The toolkit integrates modules, models and lexical resources for sentence splitting, tokenization, part of speech tagging, lemmatization, dependency parsing (Universal Dependencies) and Named entity recognition, recognizing PERSON, LOCATION, ORGANIZATION, FACILITY, and GPE (Geopolitical entity). The output of the toolkit is in conllu format and underwent appropriate conversions rendering it compatible with the ParlaMint guidelines.

3.1.4 Availability

The ParlaMint-GR corpus (v4.0) is freely available, together with all the ParlaMint corpora, through the Slovenian CLARIN node[9], and through CLARIN:EL, the Greek infrastructure for Language Resources and Technologies (v3.0, 2023)[10]. A detailed description of ParlaMint-GR is found in (Gavriilidou et al. 2023).

4. The Party Manifestos Corpus

This corpus consists of 5 sub-corpora, available through the CLARIN:EL infrastructure, collected, curated and deposited by Panteion University, member of the Greek CLARIN national network[11]. These are collections of electoral manifestos, involving programmatic stances and policy positions, as stated officially by the Greek Parliamentary Parties, in the occasions of five consecutive

[4] https://gslegal.gov.gr/?page_id=776&sort=time
[5] https://www.chesdata.eu/ches-europe
[6] https://www.hellenicparliament.gr/Vouleftes/Diateleses antes-Vouleftes-Apo-Ti-Metapolitefsi-Os-Simera/
[7] https://github.com/iMEdD-Lab/Greek_Parliament_Proceedings
[8] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-67B2-3
[9] https://www.clarin.si/info/about/
[10] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7603-8
[11] https://inventory.clarin.gr/search/party%20manifestos?repository__term=Panteion%20University%20Repository

Parliamentary elections: in 2009, 2012, January and September 2015, and 2019.

The five corpora add up to a total of approximately 1,4Mb of monolingual Greek texts, in plain txt UTF-8 format, with no annotation.

5. ParlaMint-GR Embeddings

Using the open-source fastText library and the ParlaMint-GR corpus we obtained ParlaMint-GR specific embeddings. During training, and in order to get the 100-dimensional vectors, we kept all parameters to their default values. To evaluate the quality of our embeddings we queried our model for the nearest neighbours of different words.

Table 1 shows that the 3 closest words to *Mitsotakis* (the Greek Prime Minister from 2019 till now) are *Prime Minister*, *Kyriakos* (his first name) and *Tsipras* (the previous Greek Prime Minister). Respectively, for the word *Prime Minister* the closest words are *Mitsotakis and Tsipras*. Interestingly, for the word *woman* the closest ones are *man* and *mother* and *mom*. Finally, for the word KKE, acronym of a left-wing party, the most similar words are *communist, movement*, and *comunist* (wrongly spelled).

Query word	Top 3 similar words
μητσοτάκης (mitsotakis)	πρωθυπουργός (prime minister)
	κυριάκος (kyriakos)
	τσίπρας (tsipras)
γυναίκα (woman)	άντρας (man)
	μητέρα (mother)
	μάνα (mom)
πρωθυπουργός (prime minister)	πρωθυπουργός (prime minister)
	μητσοτάκης (mitsotakis)
	τσίπρας (tsipras)
κκε (kke)	κομμουνιστικό (communist)
	κίνημα (movement)
	κομουνιστικό (comunist)

Table 1: The nearest neighbours of given words as provided by word embeddings

An additional evaluation step for the produced embeddings is to assess their ability to capture analogies between words. For this, we tested our model by seeding it with the following word triplet: *PASOK* (socialist party), *Gennimata* (president of PASOK 2015-21), and *SYRIZA* (left-wing party). Our model successfully captured the hidden analogy and returned as the most probable word the term *Tsipras,* who was indeed the president of SYRIZA.

6. Experimental Investigations of Political Ideologies

Utilizing the described datasets, we conducted a number of experiments focusing on the year 2015. This was a year of special interest for Greece, due to the political turbulence: there were 2 parliamentary elections, and also the bailout referendum, the first after many decades in the country, which was due to the financial crisis and the strict economic measures imposed by the country's creditors. Finally, this year was the first time a left-wing party (SYRIZA) came into power by forming a coalition with ANEL, a right-wing party.

6.1 Similarity of Manifestos across Parties

First, we used the party manifestos dataset and calculated the cosine similarity of the texts. Using cosine similarity to retrieve similar documents is widely used in computer science and information retrieval (Lahitani et al., 2016), (Ramya et al., 2018), (Gunawan et al. 2018). Initially, we represented each text as a vector with features the frequency of each word (bag of words). By doing this, we expect to have a quantitative measure of how similar or dissimilar the programs of the various parties are. We investigate the similarity of the following parties (Table 2): ANEL and New Democracy (ND) which are both right-wing, Golden Dawn (fascist), KKE and SYRIZA (left-wing), PASOK (socialist) and POTAMI (center-left), through their manifestos for the January 2015 elections. One quite interesting observation is that SYRIZA seems to have the lowest similarity with ANEL, the party they formed a coalition with twice during this period, i.e. despite the coalition, each party kept its ideology. Among the various explanations for this coalition, the dominant one seems to be that SYRIZA considered this coalition as the only way to form a government, despite their wide ideological differences. Cosine similarity is used to determine how similar the party's manifestos are to each other, not their ideological placement. Therefore, the fact that the similarity score of ND (right-wing) and ANEL (right-wing) is lower than the one between ND (right-wing) and KKE (left-wing) suggests that ND is using a vocabulary more similar to KKE than to ANEL in their party manifesto.

Manifesto 201501	ANEL	GD	KKE	ND	PASOK	POTAMI	SYRIZA
ANEL	1	0.821	0.751	0.781	0.8	0.833	0.776
Golden Dawn	0.821	1	0.853	0.882	0.928	0.936	0.895
KKE	0.751	0.853	1	0.84	0.856	0.887	0.824
ND	0.781	0.882	0.84	1	0.858	0.889	0.867
PASOK	0.8	0.928	0.856	0.858	1	0.96	0.903
POTAMI	0.833	0.936	0.887	0.889	0.96	1	0.882
SYRIZA	0.776	0.895	0.824	0.867	0.903	0.882	1

Table 2: Party manifestos similarity using bag-of-words

Apart from the traditional bag of words representation technique, we also computed the cosine similarity of

the manifestos using the centroids of their words' embeddings. Using this more recent and advanced representation method we aim to study if the initial results still hold or if the semantic representations of the words instead of the words themselves, give another insight.

Table 3 depicts the results. One initial observation is the very high similarity scores between all manifests. This denotes that all parties are topically very close to each other when it comes to their pre-election programmes (irrespectively of the solutions promised). Secondly, regarding the similarity between the manifestos of SYRIZA and ANEL (coalition government), we observe that now they have the second lowest similarity (with POTAMI being the most similar).

Manifesto 201501	ANEL	GD	KKE	ND	PASOK	POTAMI	SYRIZA
ANEL	1.000	0.983	0.963	0.941	0.963	0.974	0.953
Golden Dawn	0.983	1.000	0.968	0.966	0.966	0.981	0.970
KKE	0.963	0.968	1.000	0.932	0.930	0.945	0.927
ND	0.941	0.966	0.932	1.000	0.957	0.971	0.973
PASOK	0.963	0.966	0.930	0.957	1.000	0.992	0.990
POTAMI	0.974	0.981	0.945	0.971	0.992	1.000	0.988
SYRIZA	**0.953**	0.970	0.927	0.973	0.990	**0.988**	1.000

Table 3: Party manifestos similarity using word embeddings

6.2 MPs Speeches vs Party Manifestos

Aiming to compare the pre-election party manifestos with the post-election speeches in the Parliament, we placed each party manifesto on an ideological scale and studied the members of Parliament speeches against that scale. In order to achieve this, we use the unsupervised text scaling method wordfish (Slapin and Proksch 2008 & 2010), which has been widely used in political science to estimate party positions.

The objective of the investigation was to identify where the party members speeches in the Parliament are positioned compared to the party manifestos; in other words, whether speakers follow their parties' political stance(s) when addressing the Parliament. The first, most obvious observation (Figure 1), is that the Golden Dawn (GD) party is indeed placed on the far right. This is valid conceptually, as the Golden Dawn party is a neo-Nazi party with extremely racist discourse.

Worthwhile noticing is that, between the elections of January and September 2015, there was an ideological movement of the ANEL right-wing party towards the left, due to the government coalition with left-wing SYRIZA, possibly in an attempt to exhibit political homogeneity.

6.3 Gender-Related Observations

Investigating possible gender differences in ideology, we distinguished speeches made by male (M) and female (F) MPs of the right-wing ND party (see bottom lines in Figure 1). We see that their speeches (either M or F) are placed on the center-left ideological range, and certainly more to the left than their party's manifesto (Figure 1, 3rd line from the bottom). An interpretation for this might be that MPs, when delivering their speeches in Parliament, do not feel compelled to adhere to the right-wing discourse of their party, as attested in the respective manifesto.

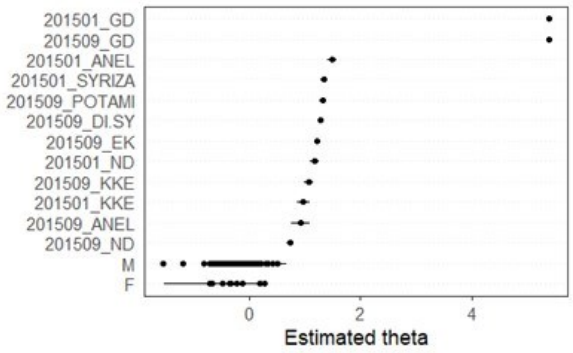

Figure 1: MPs speeches and party ideology. The Y-axis contains (from top to bottom) the parties' manifestos for the 2 elections of 2015 and at the bottom with F and M we denote the female and male MP's of ND party. The X-axis contains the ideological positions of all parties, as estimated by the wordfish method.

In the above Figure, we observe that the positioning of male and female MPs does not differ significantly, and consequently specific conclusions cannot be drawn from the specific data, concerning the position of the speeches on the ideological dimension in relation to the gender of the speaker. Since the wordfish scaling method relies on word occurrences and lexical overlap, if men and women are using similar vocabularies in their speeches, their position scores on the ideological scale will be very similar. This coincides with recent studies (Hargrave & Blumenau, 2022) reporting that the gender gap in most dimensions has narrowed in recent years.

However, it is evident from the data that the dominance of male over female MPs is still true, in numbers of MPs, and, consequently, in number of speeches. Extracting the MPs' speeches from the Parlamint-GR dataset, the descriptive statistics for the year 2015 show a clear dominance (by approximately 81%) of the Parliamentary floor by male MPs (Table 4).

Number of Speeches	
All parties	39,123
All parties Male MPs	31,604
All parties Female MPs	7,519

Table 4: Total and gender specific speech statistics for 2015

An additional snapshot of the number of speeches given by MPs of the most important political parties of

the year 2015 confirms the above observation: as shown in Table 5, female MPs are drastically less heard than their male colleagues, irrespectively of political ideology. Whether fascist (F), right-wing (R), socialist (S), or left-wing (L), women stand much less on the Parliament's podium than men: specifically, women talk 3.5 times less than men in the Greek Parliament.

Party	Gender	No speeches
ANEL (R)	F	31
	M	1011
GoldenDawn (F)	F	208
	M	1271
KKE (L)	F	321
	M	2978
ND (R)	F	1051
	M	7346
PASOK (S)	F	131
	M	1354
SYRIZA (L)	F	5363
	M	11443

Table 5: Speeches of most significant parties by gender for year 2015

7. Conclusions and Future Steps

We have presented the ParlaMint-GR corpus (its creation, the metadata used, its encoding and annotation) and the Greek party manifestos corpus. We presented some preliminary results of experiments we conducted, aiming to comparatively investigate political ideologies of parties and members of the Parliament, as expressed in these two corpora. In the immediate future we intend to broaden the scope of this study and to deal with further research questions related to political ideology as expressed in these corpora.

8. Bibliographical References

Erjavec, T. et al. (2022). The ParlaMint corpora of parliamentary proceedings. Language Resources and Evaluation. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T. et al. (2023). Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. http://hdl.handle.net/11356/1859

Gavriilidou, M., Gkoumas, D., Prokopidis P., Papavassiliou, V., and Piperidis S. (2023). The ParlaMint-GR corpus: Annotated Greek Parliamentary Proceedings, in *Proceedings of the 16th International Conference on Greek Linguistics*, 14-17 December 2023, Thessaloniki, Greece.

Gunawan, Dani, C. A. Sembiring, and Mohammad Andri Budiman. (2018). "The implementation of cosine similarity to calculate text relevance between two documents." *Journal of physics: conference series*. Vol. 978. IOP Publishing.

Hargrave, L., & Blumenau, J. (2022). No longer conforming to stereotypes? Gender, political style and parliamentary debate in the UK. *British Journal of Political Science*, *52*(4), 1584-1601.

Hjorth, F. et al. (2015). Computers, coders, and voters: Comparing automated methods for estimating party positions. In *Research & Politics* 2.2 (2015): 2053168015580476.

Lahitani, Alfirna Rizqi, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. 4th International Conference on Cyber and IT Service Management. IEEE, 2016.

Prokopidis, P, and Piperidis, S. (2020). A Neural NLP toolkit for Greek. In 11th Hellenic Conference on Artificial Intelligence (SETN 2020).

Slapin, J.B., and Proksch S-O. (2008). A scaling model for estimating time-series party positions from texts. In *American Journal of Political Science* 52.3 (2008): 705-722.

Proksch, S-O., and Slapin, J.B. (2010). Position taking in European Parliament speeches. British Journal of Political Science 40.3 (2010): 587-611.

Ramya, R. S. et al. (2018). DRDLC: discovering relevant documents using latent dirichlet allocation and cosine similarity. In *Proceedings of the 2018 VII International Conference on Network, Communication and Computing.*

9. Language Resources References

Multilingual comparable corpora of parliamentary debates ParlaMint 3.0. CLARIN:EL http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7603-8.

Party manifestos of Greek Parliamentary Parties - double Parliamentary elections 2012. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5DF1-8.

Party manifestos of Greek Parliamentary Parties - Parliamentary elections January 2015. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5DFE-B.

Party manifestos of Greek Parliamentary Parties - Parliamentary elections Sept. 2015. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5E17-E.

Party manifestos of Greek Parliamentary Parties - Parliamentary elections July 2019. CLARIN:EL http://hdl.handle.net/11500/PANTEION-0000-0000-5E26-D.