Content Reconstruction of Parliamentary Questions

Combining Metadata with an OCR Process

Fotios Fitsilis
Hellenic Parliament
Scientific Service
Athens, Greece
fitsilisf@parliament.gr

Thomas Saalfeld
University of Bamberg
Department of Political Science
Bamberg, Germany
thomas.saalfeld@uni-bamberg.de

Carsten Schwemmer
University of Bamberg
Department of Political Science
Bamberg, Germany
carsten.schwemmer@uni-bamberg.de

Abstract—**The Hellenic Parliament stores parliamentary questions using a combination of metadata extracted manually from the original text as well as the scanned document as an image file. Consequently, broad access and study of the parliamentary questions are limited as there is no principal access to the original content. A combined process was designed in order to fully reconstruct the original content of the parliamentary questions using the available metadata, which were extracted during the archivation phase, and a modified mass Optical Character Recognition (OCR) process. Post-correction of OCR results and quality controls of extracted text are paramount to ensure that the text output matches the one from the original document. The results from the OCR process are joined with the metadata and allow the full description of the original document.**

Keywords- Metadata, Optical Character Recognition (OCR), digitization, content reconstruction, parliamentary accountability, document analysis, post-correction, quality control

I. INTRODUCTION

Scrutiny of the executive is one of the fundamental functions of a Parliament. In its most basic form it takes the form of a simple parliamentary question (for simplicity: question), which a Member of Parliament (MP) may address to one or more Ministers. Such questions may serve a number of purposes: They may include requests for information, which (if provided) help to reduce informational asymmetries between the executive and Parliament; critical questions may be used by opposition parties to challenge the government on policy; where they are used to take up the concerns of citizens, backbenchers may employ them to alert Ministers to problems of policy implementation; and they may allow MPs to signal concern for the demands of citizens, as shown in [1] and [2].

In legislative studies, questions are frequently used as an indicator to assess the policy agenda of MPs and their parties [3]. In order to study parliamentary behavior, access to the content of parliamentary questions is paramount [4]. For almost a decade, the Hellenic Parliament has stored parliamentary questions as image files making these available online to scholars and the wider public. However, the image form of questions does not permit quantitative content analysis, which remains limited to a narrow set of metadata.[1]

In other European parliaments, relevant metadata and the text of questions can be retrieved digitally via a number of channels. For example, questions in the British House of Commons can be accessed via Hansard and, via an Application Programming Interface (API), through the electronic platform theyworkforyou.com. The text of questions in the Italian Chamber of Deputies can be extracted via a SPARQL RDF query interface; in the German Bundestag questions are available as PDF documents. There is no international standard, but Greek questions are still relatively difficult to retrieve for automated analysis in international comparison. In addition to academic use, the digital availability of questions enhances democratic accountability and has a potential to strengthen the link between citizens and their representatives.

In this paper we describe a combined, reproducible and reliable method to fully reproduce the original content of parliamentary questions using the metadata acquired during the archivation process as well as a modified mass OCR process using a commercially available tool. Several issues encountered during the pilot phase of the project regarding both sub-processes will be discussed and the solutions of choice will be presented.

Financial support for part of this work was given by the German Research Foundation (DFG, grant no. SA 2160/3-1) as part of the larger ORA+ project “Pathways to Power: The Political Representation of Citizens of Immigrant Origin in Seven European Democracies (PATHWAYS)”, which is co-funded by the Agence nationale de la recherche (ANR, principal investigator: Manlio Cinalli), the Deutsche Forschungsgemeinschaft (DFG, principal investigator: Thomas Saalfeld) the Economic and Social Research Council (ESRC, principal investigator: Laura Morales) and the Nederlandse Organisatie voor Wetenschappelijk (NWO, principal investigator: Jean Tillie)

[1] See the search item under “Means of Parliamentary control” at http://www.hellenicparliament.gr/en/

II. Scope of Work and Methodology

In recent years, several large-scale projects to digitize parliamentary material have been under way, as seen in [5] and [6] indicatively, and relevant studies have been conducted [7]. At the same time, the Hellenic Parliament has started several digitization projects in order to preserve older damaged and time-worn parliamentary documents.[2] This aim is also directly linked to the Open Government Partnership (OGP) initiative, where Greece has participated through a relevant action plan since 2012.[3]

As ex-ante digitization of means of parliamentary control has not been a distinct part of the Parliament's digital strategy,[4] a scientific cooperation between the Hellenic Parliament Scientific Service and the University of Bamberg allowed for the exchange of the necessary knowhow in order to initiate digitization of parliamentary questions on a pilot scale. Cooperation took place in the framework of the PATHWAYS project,[5] which studies not only the descriptive representation of citizens of immigrant origin in selected European democracies but also the way issues relating to immigration and the integration of immigrants are articulated in the latest complete legislative sessions of the national Parliaments of Belgium, France, Germany, Greece, Italy, the Netherlands, Spain and the United Kingdom (both for MPs with and without immigrant origin).

The textual data extracted from parliamentary questions are used to carry out qualitative and quantitative content analyses seeking to identify and explain cross-national, cross-party and inter-individual differences in the way questions about immigration and the integration and exclusion of immigrants are framed. Compared to an analysis of speeches or votes, questions have the advantage of being a more valid indicator of individual legislator behavior, although they can be aggregated to establish similarities and differences across parties, countries and time. Since the PATHWAYS consortium collected a complete set of questions across all policy areas, analyses on different topics can be carried out as well. In this process, the research team in Bamberg also developed a simple, Java-based online tool for the analysis of the content of parliamentary questions in Greece and the other countries under study.

The 16th legislative session (5.2.2015-28.8.2015) of the Hellenic Parliament has been chosen for the pilot application of the described processes, as it offers several advantages. First, it is reasonably short (approximately 7 months long). Hence, it contains a relatively small and manageable number of questions (4.507 questions). Second, it is a recent legislative session. This means that we could expect the scanned documents to be of high quality, due to advancements in the scanning tools, compared with older legislative sessions a decade ago. For the same reason, the manual recording of metadata from the original document at the time of submission of the parliamentary question was expected to be error-free to a high degree.

As the methods that are described here will continue to be refined in iterative steps, a concluding evaluation is still not at hand. Nevertheless, this pilot application is of particular high value due to the high data volume to be processed, which corresponds to the overall number of available digitized questions. Digitization of parliamentary questions in the Hellenic Parliament started as early as in 2006 and official data regarding their number can be found in [8]. By taking only full legislative sessions into account (12th-16th: 26.9.2007-28.8.2015 ≈ 8 years) and including the recess sessions, we come to a total of 108.428 questions, a figure that roughly corresponds to 13.500 questions per year.

III. Process Description

A. *Submission Process*

In order to describe the process of reconstruction of information contained in the original document it is important to shortly present the submission process of parliamentary questions in the Hellenic Parliament. While there are efforts to standardize the submission process electronically, it still happens in the traditional manner of presenting a hard copy of the parliamentary question to the Department for Questions. There it gets stamped and tagged with a submission date and a unique question-ID. A qualified parliamentary employee with access to the PAPYRUS Document Management System, which is built around an ORACLE Data Base, then manually extracts a set of metadata from the document and inserts them into the respective data base fields. Finally, the document is scanned, and an image PDF file is uploaded to the database.

B. *Form of a Parliamentary Question*

The content of parliamentary questions is regulated by Art. 126 of the Hellenic Parliament Standing Orders. There it is stipulated that "written questions must be clear, brief and aiming towards informing on whether an incident has actually taken place or on what measures the Government plans to adopt in order to deal with a specific issue of general or local interest". However, a special form or designated format for a question is not stipulated or suggested. In parliamentary

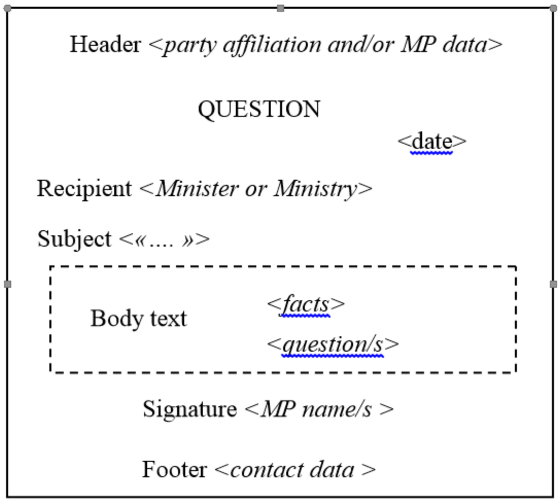


Figure 1. The form of a typical question

[2] See item 'Historical Parliamentary Archives' under 'Digital Library' at http://www.hellenicparliament.gr/en/

[3] http://www.opengovpartnership.org/

[4] http://diafaneia.hellenicparliament.gr/ElectronicDeliberation/DeliberationActionPlan2016-2018OpenGovernmentPartnership

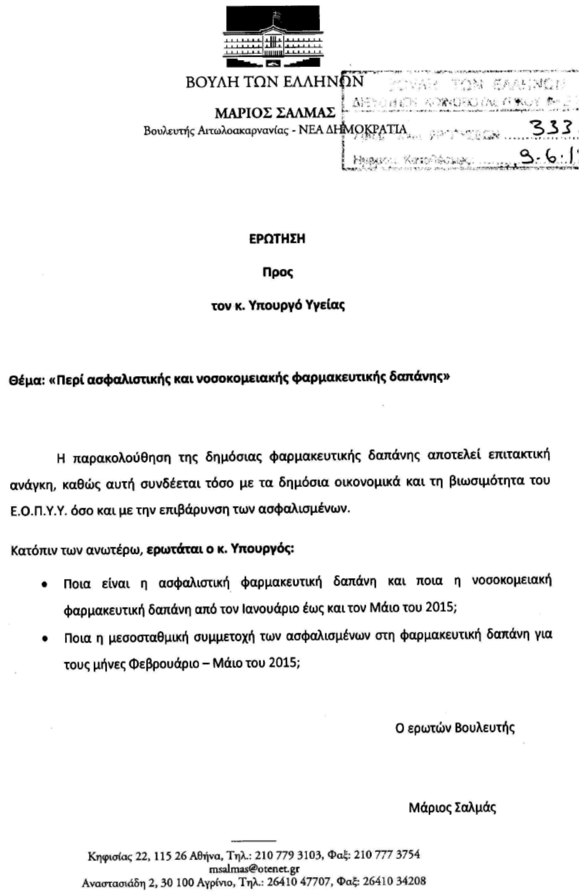
[5] http://pathways.eu/

practice, a question is not much different from a regular official letter, and MPs from all parliamentary groups follow the general schema presented in Fig.1.

According to this template, a typical question contains a header and a footer that allow for a clear identification of the parliamentarian and may include parliamentary group affiliation and office contact data. The type of parliamentary instrument, in the present case “QUESTION” (in Greek: “ΕΡΩΤΗΣΗ”), is clearly visible on the document, usually placed in all capital letters and centered. A question may also include a date. However, this is sometimes omitted as all questions are manually dated during the submission process. On the other hand, the Minister(-s) or Ministry(-ies) the question is addressed to must be clearly indicated.

The subject of the question, “ΘΕΜΑ” in Greek, immediately follows the recipient. The body text contains the argument, on which a set of questions are based. The document is signed by one or more MPs. The non-standardization of the question format has serious implications during the OCR process as will be shown in subsection D that follows.

Fig. 2 displays a parliamentary question sample (question ID: 9268302) that has all elements described above. The stamp, which is added manually during the submission process, is clearly visible and critically impairs the OCR process in its neighborhood.



ΒΟΥΛΗ ΤΩΝ ΕΛΛΗΝΩΝ

ΜΑΡΙΟΣ ΣΑΛΜΑΣ
Βουλευτής Αιτωλοακαρνανίας - ΝΕΑ ΔΗΜΟΚΡΑΤΙΑ

ΒΟΥΛΗ ΤΩΝ ΕΛΛΗΝΩΝ
ΔΙΕΥΘΥΝΣΗ ΚΟΙΝΟΒΟΥΛΕΥΤΙΚΟΥ ΕΛΕΓΧΟΥ
[illegible] 333
Ημερομ. Κατάθεσης 9-6-1

ΕΡΩΤΗΣΗ

Προς

τον κ. Υπουργό Υγείας

Θέμα: «Περί ασφαλιστικής και νοσοκομειακής φαρμακευτικής δαπάνης»

Η παρακολούθηση της δημόσιας φαρμακευτικής δαπάνης αποτελεί επιτακτική ανάγκη, καθώς αυτή συνδέεται τόσο με τα δημόσια οικονομικά και τη βιωσιμότητα του Ε.Ο.Π.Υ.Υ. όσο και με την επιβάρυνση των ασφαλισμένων.

Κατόπιν των ανωτέρω, **ερωτάται ο κ. Υπουργός:**

- Ποια είναι η ασφαλιστική φαρμακευτική δαπάνη και ποια η νοσοκομειακή φαρμακευτική δαπάνη από τον Ιανουάριο έως και τον Μάιο του 2015;
- Ποια η μεσοσταθμική συμμετοχή των ασφαλισμένων στη φαρμακευτική δαπάνη για τους μήνες Φεβρουάριο – Μάιο του 2015;

Ο ερωτών Βουλευτής

Μάριος Σαλμάς

Κηφισίας 22, 115 26 Αθήνα, Τηλ.: 210 779 3103, Φαξ: 210 777 3754
msalmas@otenet.gr
Αναστασιάδη 2, 30 100 Αγρίνιο, Τηλ.: 26410 47707, Φαξ: 26410 34208

Figure 2. Parliamentary question sample

C. *Metadata Extraction and Handling*

Metadata are expressed through Extensible Markup Language (XML) syntax using UTF-8 character encoding. The syntax of publicly available metadata from a single question is shown in Fig. 3. A total of 11 elements, field_1 to field_11, placed between the start-tag <data> and the end-tag </data>, is used to display the metadata that are made publicly available through the web page of the Hellenic Parliament. The appropriate XML data structures that are used internally by parliamentary services contain many more elements, i.e. information regarding the answers provided by the Minister(-s).

It is obvious that the most important element not recorded during the metadata manual extraction phase is the actual body text of the question. Hence, one of the core aims of the present study has been to combine the validated metadata with the textual data acquired through the OCR process. If a question is signed by multiple MPs or addressed to multiple Ministers, then fields 5, 6 and 7 (Name of MP, Name of Minister and Questioned Minister (Ministry name), respectively) contain multiple entries in form of different items.

It also has to be noted that there is a certain redundancy contained in the metadata, as seen in field_7, which contains

```
<data>
        <ID>question ID tag</ID>
        <Last-Edit-Date>Date</Last-Edit-Date>
        <field_1># of Question</field_1>
        <field_2>type of control: Question</field_2>
        <field_3>submission date</field_3>
        <field_4>subject</field_4>
        <field_5>
                <Item>
                          <ID>MP ID</ID>
                          <Name>name of MP</Name>
                </Item>
        </field_5>
        <field_6>
               <Item>
                          <Name>name of Ministry</Name>
                </Item>
        </field_6>
        <field_7>
                <Item>
                    <Name>questioned Minister (Ministry name) </Name>
               </Item>
        </field_7>
        <field_8>plenary session</field_8>
        <field_9>
               <Item>
                           <Name>parliamentary group</Name>
               </Item>
        </field_9>
        <field_10>
                 <Item>
                          <ID>file_name_ID</ID>
                          <File> file_name_ID.pdf</File>
             </Item>
        </field_10>
        <field_11></field_11>
</data>
```

Figure 3. Syntax of publicly available metadata in XML

information both on the name of the Minister and the Ministry, as part of this information appears already in field_6, which contains the name of the Ministry the question is addressed to. As questions can be signed by multiple MPs or addressed to multiple Ministers, the data is structured hierarchically, where an XML tag may refer not only to one, but to several items.

Statistical analysis often requires data to be represented in form of a rectangular matrix shape. For this reason, metadata was converted to a spreadsheet format by applying several Python scripts and regular expressions to parse the XML. During the process retracted questions were also identified and removed.

D. *End-to-end OCR Process*

The metadata captured manually does not include the question's body text. The latter is extracted by using a 10-step OCR process, which is depicted as a flow chart in Fig. 4. At the same time, the textual data acquired through the OCR process may be used for verification of already available metadata. Verification was performed manually during the pilot stage of the project and no serious mismatch could be detected between the original documents and the captured metadata. This is why it was not included as a distinct step of the OCR process. However, automated metadata verification is being considered to be applied in the full-scale application due to the large volume of digitized material and the possibility of errors during manual data entry.

The process begins by collecting the image data files (step 1). These are split in batches of 200 (step 2) before entering the OCR engine (step 3). Digitization is performed by the ABBYY FineReader® 12 professional OCR software tool. If necessary, a series of image correction operations, e.g. cropping, rotation, denoise, despeckle, dewarping etc., are performed prior to OCR (step 4), in order to increase OCR accuracy and, hence, lower the demand for time-consuming post-processing of textual data. Built-in OCR engine internal algorithms are used for the purpose.

The result of the actual OCR process is exported using UTF-8 as character encoding (step 5) and is post-corrected in part by means of the Microsoft Office Spelling Tools (step 6), similar to the process described in [9]. Finally, residual OCR artifacts and unnecessary text elements are removed (steps 7-8), in order to isolate the body text of each question, which is then tagged with its respective ID (step 9). The process is repeated until all batches have been processed and is concluded with a thorough quality control (step 10). In case of serious discrepancies, single questions or pages are being re-processed by the OCR engine (step 4 onwards).

Using the automated part of the above method which comprises steps 3-5 a throughput of approx. 1,000 questions per day could be achieved by one person using a single OCR engine. Consequently, the complete set of 4,507 questions from the mentioned legislative session could be processed within a working week. However, it needs to be noted that application of the complete process, as described in Figure 4, potentially limits the final throughput.

Thus, low image resolution and scanning quality may limit OCR accuracy. In addition, inherent increased specifications for the extracted texts (texts of parliamentary questions that will be publically available through this project cannot possibly contain errors) lead to rigorous manual quality control that may limit throughput dramatically.

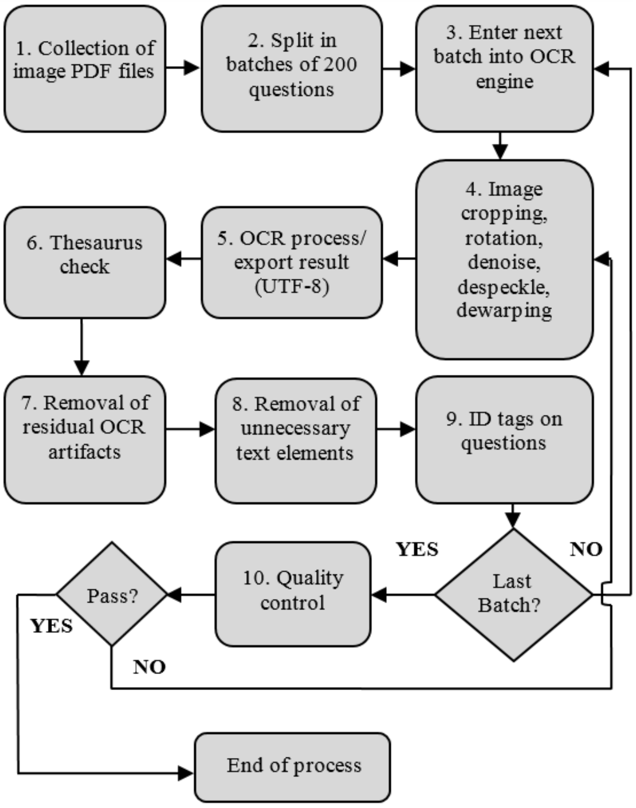

1. Collection of image PDF files
2. Split in batches of 200 questions
3. Enter next batch into OCR engine
4. Image cropping, rotation, denoise, despeckle, dewarping
5. OCR process/ export result (UTF-8)
6. Thesaurus check
7. Removal of residual OCR artifacts
8. Removal of unnecessary text elements
9. ID tags on questions
YES
NO
Last Batch?
10. Quality control
Pass?
YES
NO
End of process

Figure 4. Flow chart of a modified mass OCR process

Although there is a steep learning curve while working with this process, the bottleneck clearly lies in the quality-control step. Currently, a single person cannot exceed processing of more than 300 questions/day. As a conclusion, it would be necessary to utilize 3 persons with the proper training in parallel operation in order to match the raw speed of the automated OCR engine and avoid any procedural bottlenecks.

IV. CONTENT RECONSTRUCTION

Combining metadata with textual data from the OCR process in a last step is done by using question ID tags as identifiers for joining two data frames. In preparation for this procedure, the questions' texts were parsed via a regular expression, such that the single text file is split into one item for each question. These items and their IDs were then converted to a rectangular format and finally merged with the metadata.

This process is crucial for further meaningful analyses. Firstly, it allows researchers to combine individual questions to larger text corpora for MPs, parties or entire legislatures. This, in turn, opens up the possibility for the quantitative analysis of more general latent concepts such as ideological orientations in their dimensionality, broader topics in political debates or sentiments attached to particular political issues.

Scholars studying persistence and change in public policy will be enabled to capture the language used during the early stages of policy development, especially the emergence of an issue on the parliamentary agenda, the way policy problems are defined by different actors and the way different proposals are presented and justified.

Careful empirical studies have shown that such debates are frequently reflected in parliamentary questions and later affect the wording of the alternatives on which legislators eventually vote in the chamber. In addition, the textual data can be matched with further personal information on the individual MPs, their parties, their electoral districts, public opinion from surveys, particular policy challenges (such as budgetary, security or other crises) and time (for example the stage in the Parliament's electoral cycle).

Given the difficulties of data retrieval in the Greek case, we selected the short 2015 Hellenic Parliament with 4,506 questions tabled overall for our first validation exercise [10]. Our research strategy was to statistically identify keywords that co-occur frequently with two fundamental concepts, 'immigration' and the 'integration' of immigrants and their descendants (or functional equivalents). In a first step, we defined language- and country-specific dictionaries identifying relevant 'core' terms ('tokens') for immigration and the integration of immigrants and their descendants in each language and country. In a second step, we use these country-specific dictionaries to filter the text corpus of all questions (here: in Greece) identifying all questions that include references to tokens defined in our dictionaries. This is followed by a third step where we calculate the 'keyness' of words occurring in combination with the core tokens 'immigration', 'immigrants', 'integration' and functional equivalents [11]. In our context, keywords are terms which appear significantly more frequently in our filtered corpus than would be expected, given their relative frequency in the rest of the corpus. We determine keyness scores by applying chi²-tests with Yates correction, which are implemented in the R package quanteda [12]. Terms with the highest chi² values can then be interpreted as 'immigration' and 'integration' keywords for the corresponding country. To group terms with similar semantic meaning, we further apply 'word stemming' before keyness calculations. Stemming is a common procedure for text analysis where terms with similar semantic meaning are grouped together by algorithmically removing their prefixes and / or suffixes. In the Greek case, the results of this process can be inspected in Figure 5 where terms with the highest keyness scores are displayed. Even such a very basic procedure reveals how Greek legislators defined the problems associated with immigration in 2015.

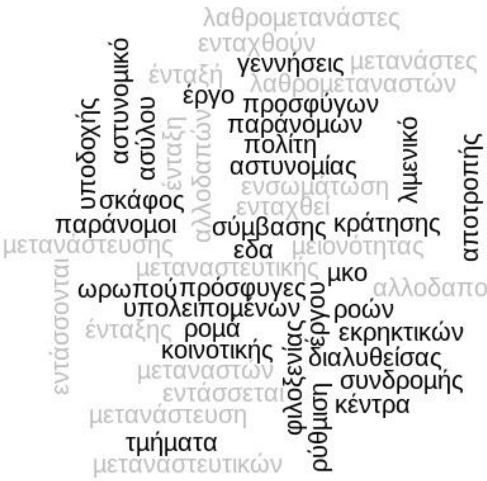


Figure 5. Word cloud from 4,506 questions in the 2015 Parliament (dictionary terms in grey, collocations with high keyness in black)

Similar to most countries in the PATHWAYS sample, the questions on immigration and integration are strongly framed by tokens relating to the 'expulsion', 'deportation', 'eviction' or 'removal' of immigrants ('αποτροπής', 'ρύθμιση'). Greek parliamentary discourse on immigration and integration is also comparable to other legislatures in highlighting the national borders (prevention ['αποτροπής'] and coastguards ['λιμενικό']) as well as aspects of security ('αστυνομικό'). However, Greek parliamentary discourse is not entirely defensive. In [10] PATHWAYS researchers found comparatively frequent references to transnational and European legal norms and obligations to grant asylum to refugees ('ασύλου', 'προσφύγων') as well as elements of a multi-cultural frame with tokens that reveal opposition to a nationalist framing of the issues, including tokens relating to 'reception' ('υποδοχής'), 'assistance' ('συνδρομής') or 'hospitality' ('φιλοξενείας') on the part of some MPs. Of course, these first impressions are very general. After expanding the sample, the metadata will thus be used to conduct, amongst others, longitudinal comparisons between MPs from different political parties, from different electoral districts and with different levels of expertise in the policy areas of immigration and integration. In addition, they will be used to identify specific topics and compare them to shifts in public opinion (measured separately via opinion polls).

V. Conclusion and Outlook

From the perspective of legislative studies in Political Science, Anthropology, Sociology, Constitutional Law, History and other disciplines, Greece is a fascinating yet under-researched case. Hardly any other member state of the European Union has undergone such dynamic and far-reaching changes to its policy agenda in the period covered here.

The nature of these changes in the EU's multi-level system has presented particular challenges to the ability of Parliament to scrutinize executive action and hold the government to account. This has brought instruments of parliamentary accountability back to the center of attention.

Furthermore, the Hellenic Parliament has included a relatively diverse ideological spectrum of MPs and parties for some time, a picture that is gradually becoming more typical of Europe overall. This motivates important further questions about the dynamics of party competition. Linked with data on public opinion in Greece, digitally available text from questions and debates in the Hellenic Parliament could reveal important patterns of elite responsiveness. Although the substantive opportunities are tremendous, the linguistic peculiarities of the Greek language and the simultaneous lack of digitally available textual data has been an obstacle to the study of parliamentary life in the Hellenic Parliament and its inclusion in major comparative analyses.

The method described in this paper allows for a full reconstruction of the content of parliamentary questions based on a proper combination of available metadata and a carefully designed OCR process. Sample results are more than encouraging and expansion to a wider scale to cover full legislative sessions is under way. It is evident that overall quality depends on the quality of metadata and textual data generated by the partial processes. The quality of metadata capturing at the moment of submission of a given question is of particular importance and steps need to be taken to unambiguously record metadata. On the other hand, a full automatic OCR engine is possible to be designed and tailored to the specifications of parliamentary questions. Ultimately, future questions need to be submitted electronically using a designated digital template and signed with the MP's digital signature.

ACKNOWLEDGMENT

The authors would like to thank the General Directorate for Parliamentary Work and the Directorate for IT and new Technologies of the Hellenic Parliament for the provision of data and the valuable technical assistance.

REFERENCES

[1] F. Russo and M. Wiberg, "Parliamentary Questioning in 17 European Parliaments: Some Steps towards Comparison," *J. Legis. Stud.*, vol.16, no. 2, pp. 215–232, May 2010, DOI=10.1080/13572334.2011.595129, [Online].

[2] T. Saalfeld, "Members of parliament and governments in Western Europe: Agency relations and problems of oversight," *EJPR*, vol. 37, no. 3, pp. 353–376, May 2000, DOI= 10.1111/1475-6765.00517, [Online]

[3] S. Martin and O. Rozenberg, The Roles and Function of Parliamentary Questions. London: Routledge, 2014

[4] M. Sánchez de Dios and M. Wiberg, "Questioning in European Parliaments," J. Legis. Stud., vol. 17, no. 3, pp. 354–367, Aug. 2011, DOI=10.1080/13572334.2011.595129, [Online].

[5] J. F. Daðason, "Post-correction of Icelandic OCR Text," M.S. thesis, University of Iceland, Reykjavik, Iceland, 2012.

[6] M. Marx and A. Schuth, "DutchParl: A corpus of parliamentary documents in Dutch," in *Proc. of the 10th Dutch-Belgian Information Retrieval Workshop*, Nijmegen, Netherlands, 2010, pp. 82-83.

[7] M. Reynaert, "Non-interactive OCR post-correction for giga-scale digitization projects," in *International Conference on Intelligent Text Processing and Computational Linguistics*, Alexander Gelbukh, Ed. Berlin-Heidelberg, Germany: Springer, 2008, pp. 617-630.

[8] F. Fitsilis and D. Koryzis, "Parliamentary Control of Governmental Actions on the Interaction with European Organs in the Hellenic Parliament and the National Assembly of Serbia," *Online Papers on Parliamentary Democracy*, vol. V, 2016. [Online]. Available: http://www.pademia.eu/wp-content/uploads/2014/02/Fitsilis_Koryzis_PademiaOnlineSeries.pdf

[9] M. Reynaert, "On OCR ground truths and OCR post-correction gold standards, tools and formats," in *Proc. of the First International Conference on Digital Access to Textual Cultural Heritage*, New York, NY, USA: ACM, 2014, pp. 159-166, DOI=10.1145/2595188.2595216, [Online].

[10] L. Geese, T. Saalfeld, C. Schwemmer and D. van der Pas, "Transnational Norms and the Substantive Representation of Citizens of Immigrant Origin," in *Pathways to Power: The Political Representation of Citizens of Immigrant Origin in Europe*. L. Morales and T. Saalfeld, Eds. Oxford, GB: Oxford University Press 2017, to be published.

[11] M. Stubbs, "Three Concepts of Keywords," in *Keyness in Texts*. Marina Bondi and Mike Scott, Eds. Amsterdam: Benjamins, 2010, pp. 21-42.

[12] K. Benoit and P. Nulty, Quanteda: Quantitative Analysis of Textual Data, An R library for managing and analyzing text. [Online]. Available: https://github.com/kbenoit/quanteda