



Development and Validation of a Corpus of Written Parliamentary Questions in the Hellenic Parliament

FOTIOS FITSILIS

GEORGE MIKROS

**Author affiliations can be found in the back matter of this article*

RESEARCH PAPER

ubiquity press

ABSTRACT

This paper presents the development of the first parliamentary corpus of written questions in the Hellenic Parliament. Moreover, we discuss a well-defined end-to-end process that has been streamlined and optimised to produce high-quality open text data based on parliamentary documents. Based on the above methodology, a representative sample of 2,000 questions from four parliamentary periods in the Hellenic Parliament has been extracted, validated, and placed into an open data repository. Furthermore, open data production is analysed, and several degrees of freedom in its application in alternative data sets are proposed and discussed. Consequently, the authors argue that this method constitutes a transferable and scalable practice that can be used by other representative institutions for the creation and subsequent study of their open data.

CORRESPONDING AUTHOR:

George Mikros

Middle Eastern Studies
Department, Hamad Bin
Khalifa University, Doha, Qatar

GMikros@hbku.edu.qa

KEYWORDS:

Hellenic Parliament;
parliamentary control; written
questions; Hellenic OCR Team;
crowdsourcing; open science

TO CITE THIS ARTICLE:

Fitsilis, F., & Mikros, G. (2021).
Development and Validation
of a Corpus of Written
Parliamentary Questions in the
Hellenic Parliament. *Journal
of Open Humanities Data*, 7:
18, pp. 1–14. DOI: [https://doi.
org/10.5334/johd.45](https://doi.org/10.5334/johd.45)

(1) CONTEXT AND MOTIVATION

The digital transformation of parliamentary institutions can be linked mainly to the availability and production of open data (Andrews & da Silva, 2013). Under certain conditions, structured open data can substantially promote parliamentary transparency and Members' accountability (Granickas, 2013). While open data production can be streamlined to become a standardised process, it is work-intensive. It puts additional pressure on parliamentary administrations due to issues related, among others, to scarce parliamentary resources, internal resistance to change, and inappropriate or absent organisational structures (Berntzen, Johannessen, Andersen, & Crusoe, 2019). Moreover, a chronic lack of consistent open data does not allow for a comprehensive understanding of parliamentary discourse. Even many large parliamentary reference corpora, such as the UK's Hansard corpus consisting of British Parliament speeches with over 1.5 billion words, offer limited analytical capabilities since they do not give access to the whole co-text because of property rights (Truan & Romary, 2020). Moreover, proprietary text formats and restricted databases reduce the annotation flexibility of the developed corpora and limit the opportunities to use the state-of-the-art linguistic annotation tools, which are mainly tuned to work with open data text formats. A poorly linguistically annotated corpus limits the research questions that can be explored and restricts the interpretability of the relevant findings introducing ambiguity, variation, uncertainty, error, and bias (Beck, Booth, El-Assady, & Butt, 2020).

To overcome these issues, an open-source crowdsourcing platform, the *Hellenic OCR Team*,¹ has been founded in 2017. It constitutes a first-of-its-kind scientific initiative for the mass processing and analysis of parliamentary textual data. It builds upon the idea that a decentralised group of people can be more than the mere sum of individuals. The team consists of a dedicated and rapidly expanding circle of experts from various sectors and disciplines, including academics, parliamentary officials, entrepreneurs, and students. Indeed, crowdsourcing has been previously utilised to build and annotate large corpora (see, e.g., Wang, Bohus, Kama, & Horvitz, 2012; Wang, Hoang, & Kan, 2013). However, this is the first reported case in which the so-called 'wisdom of the crowd' is used in a quasi-permanent volunteering format. Furthermore, a training scheme has been developed to support data validation and handling to ensure uniformity and reproducibility of the resulting textual dataset (corpus).

The Hellenic OCR Team primarily deals with the digital transformation of representative institutions. One of its main activities focuses on the compilation and analysis of a corpus of written parliamentary questions. These are questions formulated by Members of Parliament (MPs) and addressed to government members, i.e., ministers. In Westminster-type parliaments, there also exist oral questions that have been studied using corpus linguistic methods (see Zhang, Spirling, & Danescu-Niculescu-Mizil, 2017). However, in the Hellenic Parliament, questions are always submitted in written form. Moreover, parliamentary questions are the most frequently used elements of parliamentary control, whose effectiveness is a 'key concern of contemporary democratic politics' (Meinel, 2018, p. 317). This ongoing effort aims to produce and analyse a fully validated corpus of one hundred thousand (100 K) written questions. In Greece, the parliamentary control function is defined in the Standing Orders (SO) of the Hellenic Parliament (2021a), particularly in art. 124 SO. In the case of parliamentary questions, a dedicated parliamentary department within the Directorate for Parliamentary Control caters for the necessary administrative support to the process.

On the different means of control, one may consult Fitsilis and Koryzis (2016). Moreover, Fitsilis, Saalfeld, & Schwemmer (2017) offer an early outline of the corpus-building methodology for the case of the full ΙΣΤ'² parliamentary period, that is, the 16th, in the Hellenic Parliament. The current article contains a description of an enhanced and fine-tuned methodology that has been tested in the creation of a sub-corpus of 2,000 representative questions from four parliamentary periods that span over a politically turbulent decade for Greece (2009–2019). The periods included contain, among others, the handling and the aftermath of a sovereign

¹ <https://www.hellenicocrteam.gr>.

² The Hellenic Parliament uses ancient Greek numerals for numbering parliamentary periods. In the following, Greek numerals and standard numbers are going to be used interchangeably.

debt crisis and the—still ongoing—refugee/migrant crisis in Greece. The absence of relevant parliamentary open data to study and more deeply understand these issues has been a major driver behind the Hellenic OCR Team initiative.

Figure 1 exemplifies the process of breaking down a parliamentary question (link serialNr = 7122416) into data and metadata. A closer look at these parameters is given in section 3.2. Once they are extracted from the document, the original question can be fully reconstructed anytime. Section 2 discusses the development and operational framework around the Hellenic OCR Team, enabling corpus development in a reliable, flexible, and scalable manner. The tools used in the process and the underlying technology are discussed, and projections for future growth are attempted. The methodology part consists of the detailed corpus description, including sampling, artifact handling, and validation schemes (section 3). Finally, information is provided regarding the implications of this research and how it can be applied to learn more about Greece’s political and societal mechanisms during the past decade. For this, an example is included on how this corpus can help scientists analyse the political discourse around a critical social issue with international repercussions, i.e., the refugee crisis in Greece (section 4). A concluding section, section 5, provides a summary and an outlook.

The figure illustrates the extraction of metadata and body text from a Greek parliamentary question. The document is annotated with XML-like tags:

- <protocol_number>**: Points to the document number 61004/Δ1/28-5-2010.
- <date>**: Points to the date 3/6/2010.
- <submitter>**: Points to the submitter, Βουλευτής Ν. Ηρακλείου - ΝΕΑ ΔΗΜΟΚΡΑΤΙΑ.
- <party>**: Points to the political party, ΝΕΑ ΔΗΜΟΚΡΑΤΙΑ.
- <minister>**: Points to the minister, Α. Διαμαντοπούλου.
- <subject>**: Points to the subject, ΘΕΜΑ: Ανάγκη αναθεώρησης της απόφασης για την κατάργηση των εκπαιδευτικών αδειών.
- <type>**: Points to the type, ΕΡΩΤΗΣΗ.

The body text is enclosed in a dashed box and contains the following content:

Σύμφωνα με την πρόσφατα εκδοθείσα εγκύκλιος με αριθμ. 61004/Δ1/28-5-2010 εγκύκλιος του Υπουργείου Παιδείας σχετικά με τη χορήγηση και ανανέωση αδειών υπηρεσιακής εκπαίδευσης εκπαιδευτικών Π.Ε. για το 2010-2011, επιτρέπεται μόνο η ανανέωση αδειών και η χορήγηση νέων αδειών μόνο σε υποτρόφους ΙΚ.Υ. Ειδικότερα, η εγκύκλιος αναφέρει ότι: «Λαμβάνοντας υπόψη το γεγονός ότι, με την εκτέλεση του Προϋπολογισμού 2010, το σύνολο των εγγεγραμμένων πιστώσεων στον προϋπολογισμό του ΥΠΔΒΜΘ, στους Ειδικούς Φορείς 19-210 Πρωτοβάθμιας Εκπαίδευσης & 19-220 Δευτεροβάθμιας Εκπαίδευσης και στον ΚΑΕ 0282, που αφορούν στην αποζημίωση εκπαιδευτικών που τελούν σε άδεια υπηρεσιακής εκπαίδευσης μεταφέρθηκε στους προϋπολογισμούς των Διευθύνσεων Εκπαίδευσης για την καταβολή των ανεξόφλητων υποχρεώσεων οικονομικού έτους 2009, για το σχολικό έτος 2010-2011, επιτρέπεται μόνο η ανανέωση αδειών υπηρεσιακής εκπαίδευσης».

Ως εκ τούτου, κανένας εν ενεργεία εκπαιδευτικός δεν δύναται να πάρει, για πρώτη φορά, εκπαιδευτική άδεια για μεταπτυχιακές σπουδές το 2010-2011, με μοναδική εξαίρεση τις υποτροφίες του ΙΚ.Υ., οι οποίες βεβαίως χορηγούνται υποχρεωτικά σύμφωνα με τον Δημοσιοϋπαλληλικό Κώδικα. Γεγονός είναι ότι η εν λόγω απαγορευτική διάταξη αποτελεί ουσιαστικό εμπόδιο για την επιστημονική αναβάθμιση των εκπαιδευτικών και της ποιότητας της εκπαίδευσης, ενώ έρχεται σε πλήρη αντίθεση με τις εξαγγελίες της κυβέρνησης και της πολιτικής ηγεσίας του Υπουργείου περί αναβάθμισης της δημόσιας εκπαίδευσης, αλλά και με τις διατάξεις του Δημοσιοϋπαλληλικού Κώδικα που δίνουν τη δυνατότητα λήψης εκπαιδευτικών αδειών.

Δεδομένου ότι η κατάργηση της επιμόρφωσης των εκπαιδευτικών υποβαθμίζει ουσιαστικά την ποιότητα της παρεχόμενης εκπαίδευσης και της κατάρτισης των εκπαιδευτικών,

ΕΡΩΤΑΤΑΙ Η ΑΡΜΟΔΙΑ ΥΠΟΥΡΓΟΣ

Προτίθεται η πολιτική ηγεσία του Υπουργείου να αναθεωρήσει την ομολογουμένως λανθασμένη στάση της και να προβεί στην άμεση ανάκληση της ισχύος της εγκυκλίου με αριθμ. 61004/Δ1/28-5-2010 σχετικά με τη χορήγηση και ανανέωση αδειών υπηρεσιακής εκπαίδευσης εκπαιδευτικών.

Ο ερωτών Βουλευτής
Λευτέρης Αυγενάκης
Βουλευτής Ν. Ηρακλείου

body text: 7122416.txt
signature (redundant information)

Figure 1 An example of question data (body text) and metadata (marked as XML elements).

(2.1) THE HELLENIC OCR TEAM

The Hellenic OCR Team (or, simply, Team) aims exclusively to process and study parliamentary texts. So naturally, the application, handling, and further development of OCR and data validation processes stand at the core of its endeavours. New members receive initial basic training on entering the group, while more experienced members, called ‘mentors’, provide peer-to-peer advice and support. They join an international scientific network and get direct access to scientific projects while having the opportunity to acquire valuable new skills and hands-on work experience with state-of-the-art tools and methods in the greater area of digital humanities. Scientific work follows a Proof-of-Concept (PoC) approach. The Team is guided by a steering committee and is organised into four permanent groups, i.e., the *OCR group*, the *analytics group*, the *development group*, and the *parliamentary development group*. Whenever necessary, sub-groups are established on-demand to tackle new PoCs, which can also include non-members.

The Hellenic OCR Team is a non-profit initiative. Involvement with the team is only considered on a voluntary basis,³ though it has been observed that individual members have capitalised on their experience and knowledge to be later employed by institutional members. Individual members are permanent (not on a project-based basis), while there are also opportunities for institutional membership. As of April 2021, there are 39 individuals and four institutional members active. The individual members constitute the pool of experts and form an international expert network. A survey in late March 2021 among the entire population of the individual team members (N = 39) captured the Hellenic OCR Team’s demographics and other essential characteristics.

Figure 2 shows the development of the member population (active personal members vs. admission date at the time of issuing the membership certificate). It is evident that since the Team’s foundation in November 2017, admissions follow a positive linear trend ($R^2 = 0.919$).

Gender distribution shows 59% male (23) and 41% female (16) members from 13 different countries spanning four continents, Europe, Asia, North America, and South America, as visualised in the map of **Figure 3**. Unsurprisingly, a total of 35 members out of 39 are Greeks, while there are also native members from Argentina, Italy, and Cyprus, thus providing an international dimension to the initiative. Almost two-thirds of the members (25 from 39, or 64.1%) operate from Greece. **Figure 4** depicts the working sectors of the Team’s members. It turns out that the Team is heavily relying on the private sector, which contributes more than half of the members, and academia, which roughly contributes one-third of the members. This distribution is not a coincidence but the result of a long-term strategy to support and possibly enhance the capacity of legislatures with scientific know-how and technical expertise.

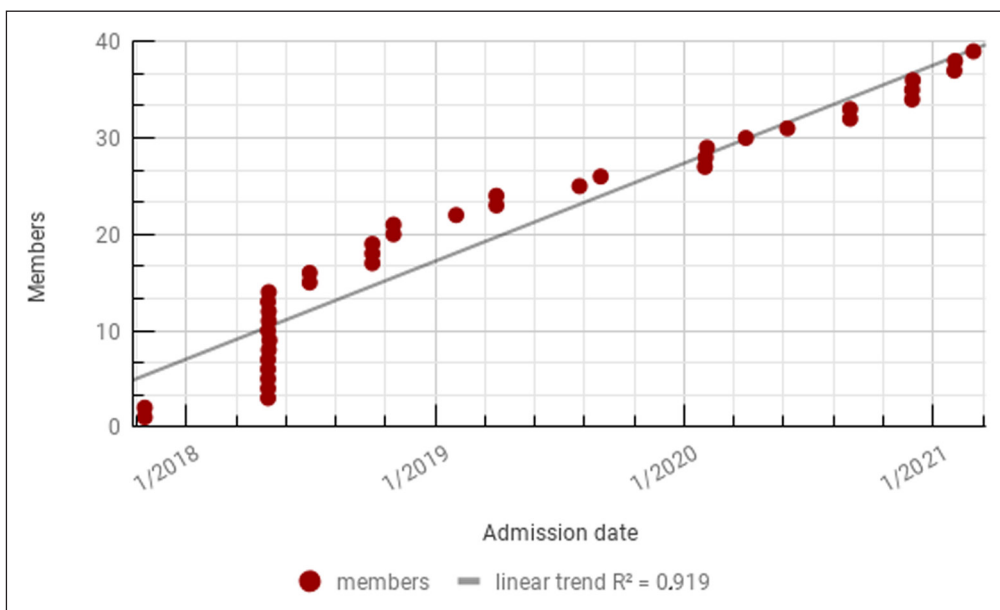


Figure 2 Development of member population.

³ No subscriptions or other side costs are linked to the membership.

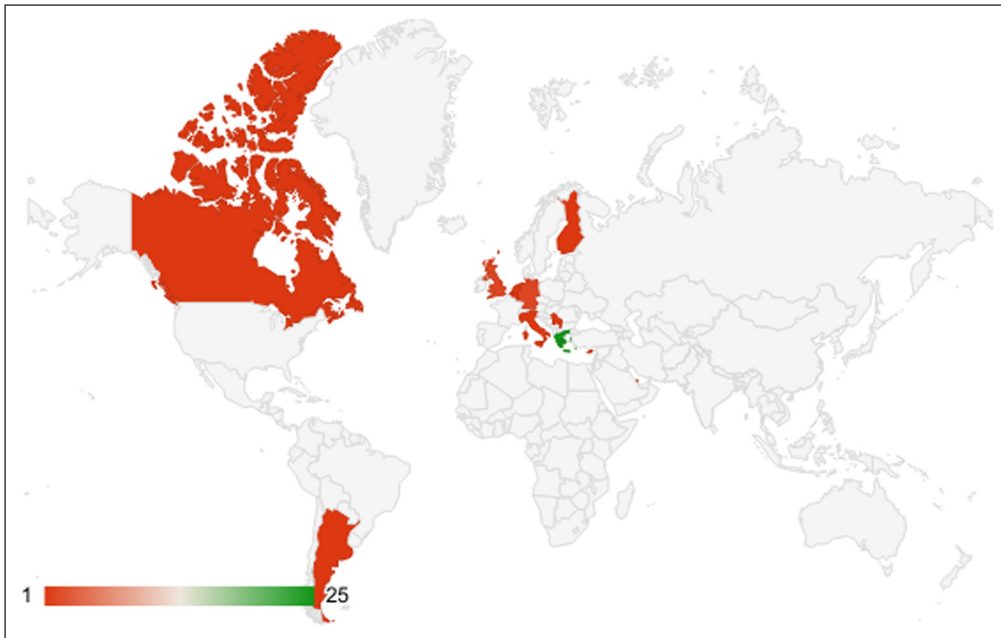


Figure 3 Geographic distribution.

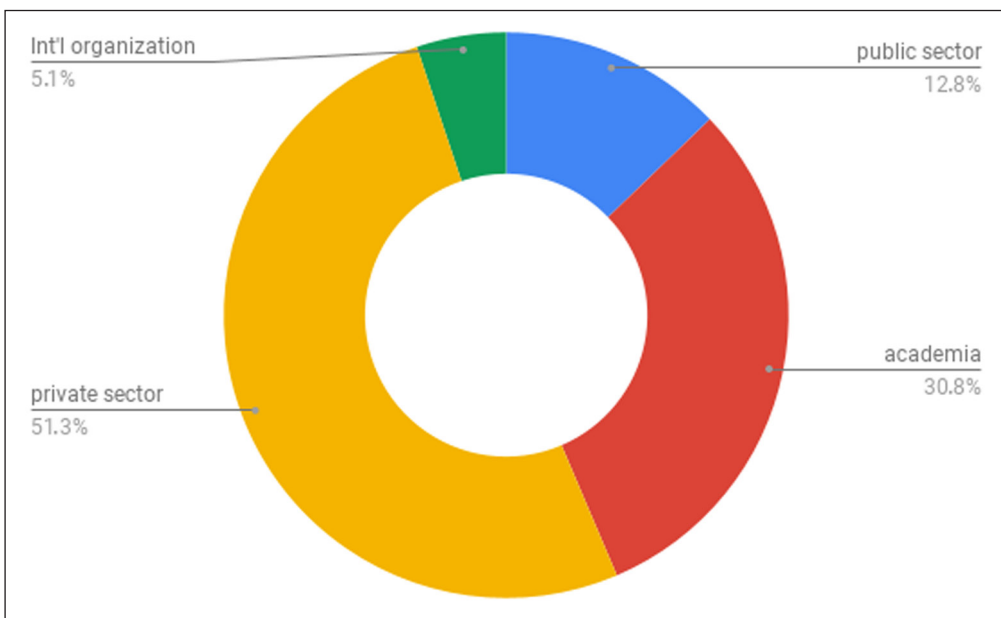


Figure 4 Working sectors (basic assumptions: students are attributed to academia and European Union institutions to international organisations).

The principal academic background of the members is shown in **Figure 5**. The Philology faculties have supplied more than one-third of the members (35.9%). In the early days of the Hellenic OCR Team, most of these members started as graduate students and now continue to contribute as post-graduate or doctoral students and professionals. This can be explained by their interaction with novel corpora from the generally under-researched parliamentary workspace that sparks excitement among the linguistic community. Most of the linguists populate the OCR and analytics groups mentioned above. Roughly a quarter of the members (25.6%) have an academic background in engineering or informatics.

From the last-mentioned members comes most of the technical expertise that the public sector lacks when developing open datasets and open-source software for their analysis. Several of these members form the development group responsible for designing the Team's software tools and solutions. Another quarter has a political or social sciences background. These members are mainly concerned with the socio-political research questions that can be formulated and empirically validated in the developed corpus. They largely populate the parliamentary diplomacy group but also take on other horizontal tasks. Finally, there are four lawyers (10.3%) and one health science professional interested in various specialised PoCs.

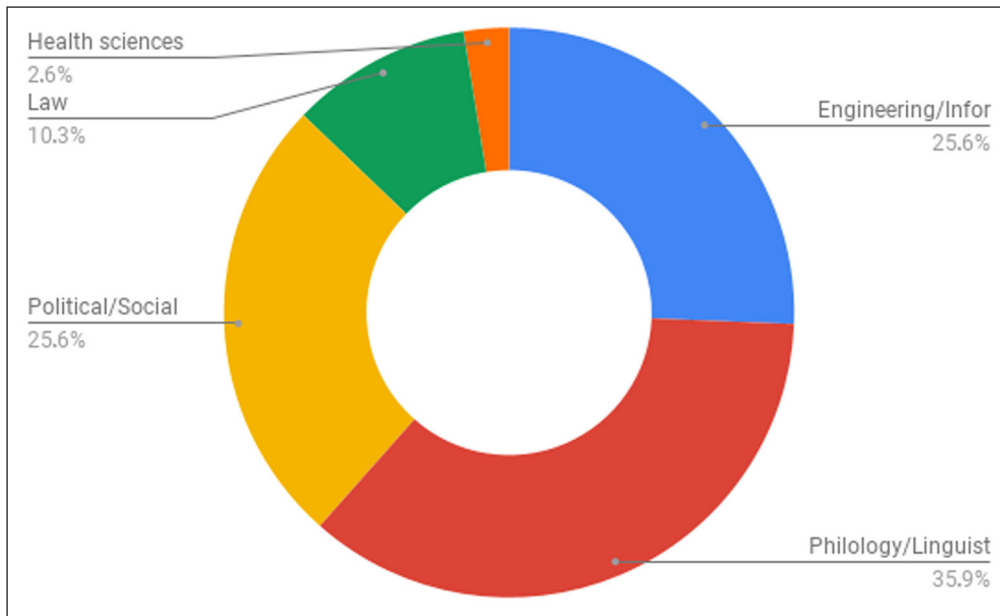


Figure 5 Academic background.

The average time members dedicate to the Team can vary greatly. Almost half of them (48.7%) have indicated that they invest less than an hour per week (hpw). Roughly a third (35.9%) spends 2–4 hpw in the Team’s activities, while 15.4% offer more than 4 hpw of the time.

(2.2) PROCESSING AND ANALYSIS TOOLS

The Hellenic Parliament has developed a dedicated IT structure to store and distribute documents both internally and through its website to the public. Written parliamentary questions are stored in the internal document management system (DMS). They are accessible as image PDFs via a graphical user interface (GUI) at a dedicated position of the parliament’s website (Hellenic Parliament, 2021b). This is where crowdsourcing steps in. After the data collection step, OCR converts image PDFs to documents in simple text (txt) format. Next, team members process large text units, referred to as ‘packages’, assigned to them, including the body text of written questions. Processed packages pass through a final quality control step, and the corpora are linked to their respective metadata, also available at the same position on the mentioned website. The resulting dataset then populates a database while pipelined for scientific exploitation. The detailed process has been defined by Fitsilis et al. (2017), and a slightly modified process used for capturing the corpus presented in this article is shown in the next section. As the metadata and body text of questions are captured in different processes, they are subsequently combined to fully reconstruct the original content.

Figure 6 shows the modified process for handling a single batch of files (package). The graph shows the four major tasks towards the production of open parliamentary data as distinct processing steps, A to D. The corresponding digital tools, custom made or proprietary, necessary for each processing step are also indicated. This is a high-level, simplified representation as there can be several iterative actions within a single step. For instance, OCR optimisation tests

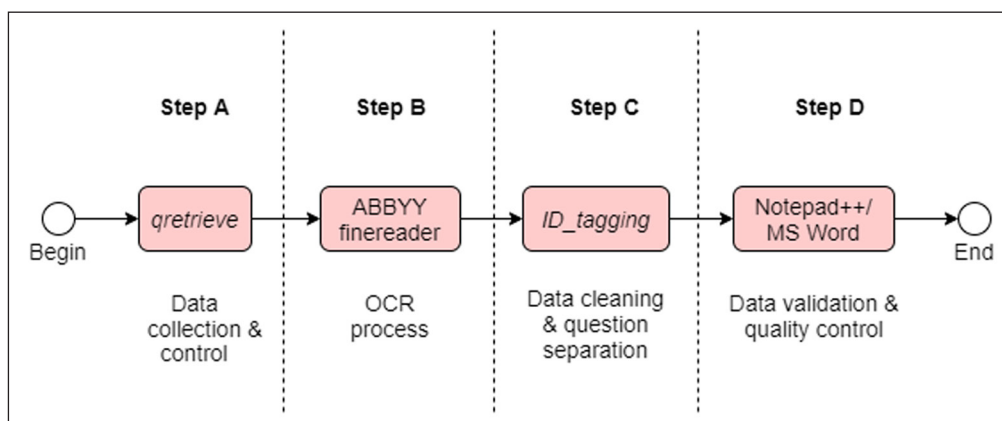


Figure 6 Modified processing steps and tools.

might be needed for single files due to the low image quality of the original PDF document. Similarly, the naming/tagging process for question/body-text separation might require several attempts. Apart from data consistency control during the first step, steps A to C are mostly automated. The human operator provides the input parameters for executing the custom software described below and operates the OCR tool through an advanced GUI. The concluding step (D) is performed manually using text processing tools with a thesaurus add-on. Process optimisation through extensive testing, fine-tuning, and a steep learning curve due to efficient member training has significantly shortened processing time. Indicatively, for a single package, end-to-end processing would typically take a full working day, with most of the time consumed by the data validation and quality control step.

A set of open-source R and Python scripts is used for data collection, processing, and parsing. The processing applied to build the present corpus includes a web scraper (*qretrieve*) and an indexer (*ID_tagging*) (both coded in the R programming language) to extract the sampled documents with their metadata and to create the bulk text (the result of the OCR process) with the respective questions.⁴ An additional tool, *Xtralingua*,⁵ has been developed to extract quantitative text profiles from multilingual corpora (Fitsilis, Leventis, Mikros, & Papantonakis, 2020). The *qretrieve* script is used to extract both metadata and question files from the parliamentary portal.⁶ Since a respective native application programming interface (API) to extract the envisaged corpus of questions is missing, *qretrieve* screens the portal's source code that displays the questions using a common visual pattern. In the underlying matrix, the script identifies a set of pre-recorded elements (metadata parameters) per question, including the links to the actual question and answer files, which are then captured on a csv file. An additional step is necessary to download the linked question files.⁷ Upon availability of the image PDF documents, OCR batch processing produces a single text file with no identification of individual questions. In order to separate questions into dedicated text files, the body text of the questions is first manually cleared from unnecessary elements and OCR artifacts, resulting in unified text blocks separated by line breaks. Starting from the top of the text file, which is tagged using the universal registry number of the first question in the batch, the *ID_tagging* script identifies consecutive line breaks, interprets them as the beginning of a new question's body, and tags them (<ID> universal registry number </ID>) with the number of the question next in line.

(3) METHODOLOGY

(3.1) CORPUS DEFINITION

A parliamentary period is defined as the time interval between two consecutive parliamentary elections. The corpus presented here consists of four equally large samples of written parliamentary questions (art. 126 SO) taken from four Hellenic quasi-consecutive parliamentary periods, including their summer recess sessions. The selected periods were the latest four fully concluded ones, i.e., from II' (13th) until IZ' (17th) with almost 100,000 written questions. The IA' parliamentary period (14th) has not been included as it has been a brief transitional period during which no parliamentary control took place. It should be noted that under the written questions (Q), we also include the ones that contain an Application for Submission of Documents [English translation for *Aitisi Katáthesis Engráfon* (AKE)] (art. 133 SO).

The corpus creation process follows the established Hellenic OCR Team methodology with three significant differences. First, all PDF files and their related metadata are extracted using an R-based web scraper. Second, the package size for the batch OCR process and subsequent validation has been increased from 200 to 500 files because of the sharp learning curve on the member's side. Third, following data collection, document quality is controlled to ensure that the data packages are error-free (see section 3.4).

⁴ The *qretrieve* and *ID_tagging* scripts are available under GPL-3.0 licence. Source code to be retrieved from <https://github.com/hocrt/qretrieve> and <https://github.com/hocrt/Rscripts>, respectively.

⁵ This tool undergoes further development to boost research in the fields of computational stylistics, text mining and beyond. *Xtralingua* is available under MIT licence.

⁶ The links to the PDF answer files are also logged in the metadata. Yet, answers have not been part of this project and, hence, they are not included in the corpus.

⁷ This step involves the use of the cURL utility.

(3.2) CORPUS METADATA

While the body text of questions is contained in corresponding text files named after their serial number, their metadata is placed in a single labelled csv matrix. Each question entry occupies a row and is characterised by a series of metadata placed in columns. Metadata parameters allow for full characterisation and the complete reconstruction of the original question. These include a universal and a local registry number, called *link serialNr* and *protocol number*, respectively.⁸ The universal number matches the name of the question text file, thus allowing linkage between textual data and metadata. To specify the exact parliamentary session and period the question was asked, the parameter *session/period* is used. Parameters *date* and *date last modified* are self-explanatory, while *type* describes the nature of the parliamentary control means. Here, it takes two different values, *erotiseis* (questions) and *erotisi se syndyasmó me AKE* (question combined with an application for submission of documents). Next, a general description of the question, usually a one-liner, is included in the parameter *subject*. It is followed by the parameters *submitter* and *party* containing the name of the MP who submits the question and the parliamentary group the Member belongs to. There are also parameters to identify the addressees, that is, the *ministry* and the respective *minister*.⁹ Ultimately, actual links to the original question and answer files (these are the non-processed image PDF files) are contained in the respective metadata parameters.¹⁰

(3.3) SAMPLING

Transforming 100,000 questions into open data is a challenging task since, even with an intensive crowdsourcing methodology, the entire corpus processing could easily yield to a year-long project. Therefore, sampling is necessary to be able to perform a baseline annotation and some initial exploratory linguistic analysis. The identity of the samples is shown in [Table 1](#).

Q/Q+AKE (PERIOD)	PERIOD FROM-TILL (MONTH/YEAR)	TOTAL NUMBER OF Q (N)	SAMPLE SIZE (S)	PERCENTAGE OF THE TOTAL NUMBER OF Q THAT THE SAMPLE REPRESENTS (%)
IZ' (17th)	10/2015–6/2019	32,910	500	1.5
IS'T' (16th)	2/2015–8/2015	4,605	500	10.9
IE' (15th)	6/2012–12/2014	27,377	500	1.8
II' (13th)	10/2009–4/2012	35,103	500	1.4
Total	10/2009–6/2019	99,995	2,000	2.0

Table 1 Sampling overview.

For methodological reasons, the package size has been chosen to match the team's standard operating procedure. Hence, a sample of 500 questions per period has been randomly extracted. As for the sampling strategy, representative samples were randomly extracted using the *qretrieve* script applying the *sample()* R function. Moreover, the sample size needed to be balanced so that manual correction could be feasible under a specific timeframe and, at the same time, the end product could reliably represent both the linguistic and the thematic content of each sampled period. The size of the compiled corpus was 638,865 tokens and 43,025 types. The detailed size statistics per period and some basic stylometric indices can be found in [Table 2](#).

It is evident that, linguistically speaking, the corpus compiled is a highly homogeneous resource. Standardised Type/Token ratio (TTR), which roughly accounts for the vocabulary diversity, appears to be very consistent with small variations among the different periods. Since TTR is highly sensitive to text size, the calculation has been normalised to a standard text-size chunk

⁸ The use of a second (local) registry number is deemed necessary for numbering and identifying questions within individual parliamentary periods.

⁹ Single questions can be signed by more than one MPs of the same party. Questions can be addressed to more than one ministry (minister) too.

¹⁰ It is worth noting that the parameters *type*, *subject*, *minister* and/or *ministry*, *submitter* and *party* are defined by the MP who drafts and submits the question. The rest of the metadata are added by the parliamentary administrators for the purpose of enhanced document handling and archiving.

TEXT FILE(S)	TOKENS	TYPES	STANDARDISED TTR	MEAN WORD LENGTH (IN CHARACTERS)
IZ sample data final.txt	162,874	21,409	50.10	5.71
IΣT sample text final.txt	181,147	21,668	49.08	5.67
IE sample final.txt	173,741	21,378	49.35	5.65
IF sample final.txt	121,103	16,238	47.16	5.65
All four periods	638,865	43,025	49.05	5.67

Table 2 The corpus size and some basic stylometric indices.

of 1,000 words (Bucks, Singh, Cuerden, & Wilcock, 2000). Furthermore, a second stylometric feature, the mean word length, is also highly stable across the various periods, which further corroborates our corpus homogeneity.

(3.4) QUALITY CONTROL AND VALIDATION

After data extraction and before performing OCR, the PDF files and their respective metadata were thoroughly controlled for potential errors such as missing or corrupted files and data inconsistencies. These contain, for instance, possible revisions or removal of questions that left artifacts in the database and data entries without accompanying documentation. The percentage of encountering such issues during sampling ranges between 0.5% and 2.3% (1.2% on average). To maintain a sample size of valid 500 questions, corrupted elements were substituted by random ones from the respective large corpus of questions.

After the OCR, a validation step is performed on the textual data. Because of the critical nature of the documents included in the corpus, supervised/unsupervised NLP methods for OCR post-correction and content normalisation were ruled out (see, e.g., Reynaert, 2014; De Clercq, Schulz, Desmet, Lefever, & Hoste, 2013). Instead, manual correction and corpus clean-up were preferred. In addition, standard training catered for normalisation of contents and data handling, and internal user guidelines were created to homogenise the validation process (see section 3.5). Corpus validation was necessary because OCR algorithms were underperforming frequently in our data. Moreover, there was no standard question format available, which reduced the quality of textual representation of the original question documents.

In order to determine OCR quality, the Low Confidence Characters (LCC) index was used (for dealing with LCC, see, e.g., Le, Straughan, & Thoma, 2002; Castiglia & Walter, 2008). For this, a random 500 questions sample from the 15th period was studied. OCR of question documents results in a body of raw data on which data validation is being conducted. Validation is done in two major steps, *question body isolation* (QBI) and *removal of residual artifacts* (RRA). The development of the LLC index (expressed in percentage terms) throughout the process can be viewed in [Table 3](#).¹¹

STEP	QBI	RRA	QUALITY CONTROL
LCC index	97.2%	98.3%	≈100%

Table 3 Development of the LLC index.

(3.5) ARTIFACT HANDLING RULES AND FORMAT

Handling of OCR artifacts follows empirically developed guidelines. During training, members need to process under the supervision of mentors at least two data packages, i.e., 1,000 questions, before they are allowed to process any packages independently. A concluding quality control step included inter-annotator agreement scores¹² and checks for any discrepancies in artifact handling. In general, rules of conduct prohibit correction of grammatical or syntax

¹¹ The index values were calculated based on the relevant internal feature of the ABBYY FineReader 14 application. For the case of the QBI, the result was reproduced by the ABBYY FineReader engine 11, which calculated 2.68% ‘suspicious characters’, i.e., 37,603 of 1,405,026.

¹² The inter-annotation agreement was calculated on a standard package of 500 questions, with two human annotators correcting the raw OCR output. Evaluation on the same package was repeated with three pairs of annotators and the average Cohen’s κ was 0.94.

errors unless these appear as the result of the OCR process. The editor does not interfere with spelling issues, except in extraordinary cases such as obvious anagrams. Particular care has been given to geographical named-entity normalisation since locally, or regionally used spelling might deviate from the widely used form.

Specific anonymisation rules were applied in order to protect personal data (e.g., names, personal identification numbers, and contact details). These were meticulously removed from the body text and replaced by empty brackets, e.g., {...} or [...]. Personal data protection did not apply to public figures or officials. MPs occasionally attach supportive graphical and textual material and, on some occasions, earlier versions of the same or/and other written parliamentary questions. These add-ons are not considered to be part of the body of the questions. When handling two-dimensional tables, commas ‘,’ are used as column separators. Footnotes are moved to the end of the question body with the footnote number in brackets. In the chosen format, split words are necessarily put together. Finally, the subject line, as well as any greeting form, is removed.

The total corpus (2,000 text files) and the metadata file (csv format) have been uploaded to Zenodo, a general-purpose open-access repository for research data. The corpus is distributed under the open-access license Creative Commons, Attribution-NonCommercial 2.0 Generic (CC BY-NC 2.0). **Table 4** summarises some of the most significant corpus attributes such as text format and language, licensing and important dates, and points at the permanent repository where the dataset rests.

ATTRIBUTE	VALUE
Text format	Plain text (txt)
Encoding	UTF-8
Data format	<ul style="list-style-type: none"> • Corpus metadata: csv file • Corpus files: txt files
Creation date	April–June 2019
Publication date	10 May 2021
Language	Greek
Licence	CC BY-NC 2.0
Repository	Zenodo: https://zenodo.org/record/4748989
DOI	10.5281/zenodo.4748989

Table 4 Corpus attributes.

(4) IMPLICATIONS AND APPLICATIONS: A SMALL CASE STUDY ON THE SEMANTIC PROSODIES OF THE GREEK TERMS ‘REFUGEE’ AND ‘MIGRANT’ IN THE PRESENTED CORPUS

The possible research questions that can be addressed are many and diverse. Two main research directions are envisaged to be served by the exploitation of this corpus:

- a) Linguistic research: Although there is some research on the Greek political discourse, it is fragmented and constrained by the limited availability of relevant open data. The described corpus will offer the first comprehensive, systematic, and detailed collection of texts related to the parliamentary control procedures of the Hellenic Parliament. The corpus organisation in parliamentary periods helps researchers form time-sensitive research questions and watch how specific linguistic features, semantic spaces, and words with significant sentimental or ideological value change over time. Moreover, the time frame of the corpus coincides with the most turbulent time in the economic history of Modern Greece since it covers the period in which Greece was under severe economic pressure and international financial restrictions. Parliamentary questions encode the dynamics of the socio-political environment and allow us to study its impact on the linguistic messages produced. They also provide valuable insights into how different political parties approach sensitive socio-economic and international issues using either alternative words or the same terms with different collocations and semantic prosodies.

Furthermore, since the described digital content is available in an open and structured format, it enables the use of novel tools and methods from the fields of computational linguistics and artificial intelligence (Kouklakis, Mikros, Markopoulos, & Koutsis, 2007; Markopoulos, Mikros, Iliadi, & Lontos, 2015), offering a testbed for further development of these technologies for the Greek language.

- b) Socio-political research: The availability of unified and verified corpora like the described one allows for interlinking several -formerly distant- areas of research, e.g., history, political science, social psychology, and others, thus opening new horizons in the understanding of parliamentary information and discourse. Various research questions can be formulated about the corpus, including how critical foreign affairs issues are formulated and what concepts are most relevant for the Greek foreign policy. Moreover, how specific critical socio-political issues like the refugee crisis are being framed in the broader public debate and its social consequences. The application of advanced text mining methods like sentiment analysis and network analysis can also help answer complex questions linking specific politicians, political parties' affiliations, and their position on current critical legislature issues. This kind of layered approach can uncover deeply rooted links between individual actions, ideologies, and social interaction, offering plausible interpretative models of political action and public engagement.

To showcase the usefulness of this corpus to actual research related to both linguistic and sociopolitical research questions, we will present some preliminary findings from a larger study related to the Greek terms for 'migrant' and 'refugee' (μετανάστης [*metanástis*] and πρόσφυγας [*prósfygas*], correspondingly) (Giovani, Krimpas, Fitsilis & Mikros, accepted). The Greek terms 'refugee' and 'migrant' are very different in terms of semantic content, grammatical relations, and collocation networks. The European Commission's department in charge of migration and home affairs (DG HOME) defines 'migrant' and 'refugee' as follows:

- Migrant: In the global context, a person who is outside the territory of the State of which they are nationals or citizens and who has resided in a foreign country for more than one year irrespective of the causes, voluntary or involuntary, and the means, regular or irregular, used to migrate.
- Refugee: In the global context, either a person who, owing to a well-founded fear of persecution for reasons of race, religion, nationality, political opinion or membership of a particular social group, is outside the country of nationality and is unable or, owing to such fear, is unwilling to avail themselves of the protection of that country, or a stateless person, who, being outside of the country of former habitual residence for the same reasons as mentioned before, is unable or, owing to such fear, unwilling to return to it.

It is evident from the above definitions that these two terms represent two different senses and should be used in different discourse contexts, especially in the framework of political communication. However, a first corpus linguistics analysis of the two terms in the described corpus offers a different perspective. It seems that in the discussions of the refugee crisis in Greece, politicians use these two terms interchangeably, converting them to synonyms. One of the most striking observations was that although the term 'migrant' does not have any negative semantic quality in its definition, it gets an intense negative semantic prosody in Greek Parliamentary discourse. The five most frequent collocates of 'migrant' are all words with negative meaning. More specifically, the most frequent adjective before 'migrant' was παράνομος 'illegal' with 111 occurrences (71% of all occurrences of the Adj + 'migrant'), followed by παράτυπος 'irregular', οικονομικός 'economic', μη νόμιμος 'non-legal', επίδοξος 'intended'. All these modifiers are focusing on the illegal status of the migrants in Greece and create negative connotations when the term 'migrant' is used in Greek.

On the other hand, the term 'refugee' has a more neutral semantic prosody than the term 'migrant' in the Greek parliamentary discourse. The five most frequent collocates of 'refugee' in the corpus are a) τηλεφωνήματα 'phone calls', b) φιλοξενία 'hosting', c) μεταφορά 'transportation', d) ταυτοποίηση 'identification', e) ένταξη 'integration'. It is evident that the term 'refugee' is not linked with any negative term, and it is used mostly in contexts related to the hosting status and the management of these people. The two terms ('migrant' and 'refugee') co-occur very frequently (54 times or 64% of every conjunctive structure that contain either 'migrant' or 'refugee' in the corpus). This high co-occurrence rate makes 'refugee' being used as a synonym

of ‘migrant’, inheriting its negative collocations. This kind of sense merging in Greek political discourse related to the refugee crisis raises broader concerns regarding social acceptance and perception of this acute social issue. The described corpus can assist us in detecting this kind of semantic biases in the political discourse and raising awareness of sensitive issues that should be reframed in a more inclusive and socially sensitive discussion.

(5) CONCLUSION

The linguistic peculiarities of the language combined with the lack of open data on parliamentary control raise significant obstacles in the study of parliamentary life in Greece. This article described a crowdsourced corpus-building process and the publication of a representative sample of 2,000 written parliamentary questions. The presented corpus demonstrates that parliamentary research can be successfully integrated into a broader scholarly enterprise based on the development of structured linguistic resources. It enriches a dynamic pool of parliamentary corpora, which is mainly developed around the ParlaMint CLARIN initiative (Erjavec et al., 2020), enhancing the ability of scholars to identify and document parliamentary discourse in a cross-linguistic context. The efforts herein showcase both the strengths of crowdsourcing and the value existing behind decentralised networks of experts. The authors envisage their paradigm to be followed by other parliaments offering high-quality corpora as part of their standard task cycle and releasing them as open data in the international research community. Democracies should be based on full transparency. Sharing the full linguistic production of our governmental operations, we will then be better equipped to deal with the challenges of a changing society in the wake of globalisation.

ACKNOWLEDGEMENTS

The authors would like to thank the Hellenic OCR Team, particularly the members who contributed to corpus development and validation, i.e., Evangelos Dimiou, Alexandros Fikas, Alexandra Fiotaki, Foteini Kakaitza, Anna Karampali, Vassilis Kesidis, Marina Kousta, Sotiris Kranias, Polytimi Mountanea, Anna-Maria Moutsai, Spyridoula-Anna Pappa, Nasia Pliakogianni, Natasa Theochari, Katerina Tzortzi, and Eleni Zisioglou. The best is yet to come!

COMPETING INTERESTS


The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Conceptualisation: F.F. and G.M.; Methodology: G.M.; Validation: F.F.; Data Curation: F.F.; Writing – Original Draft: F.F. and G.M.; Writing Review & Editing: F.F. and G.M.; Supervision: F.F.

AUTHOR AFFILIATIONS

Fotios Fitsilis  orcid.org/0000-0003-1531-4128
Scientific Service, Hellenic Parliament, Athens, Greece

George Mikros  orcid.org/0000-0002-4093-5973
Middle Eastern Studies Department, Hamad Bin Khalifa University, Doha, Qatar

REFERENCES

- Andrews, P., & da Silva, F. S. C.** (2013). Using parliamentary open data to improve participation. In *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance* (pp. 242–249). NYC: Association for Computing Machinery (ACM). DOI: <https://doi.org/10.1145/2591888.2591933>
- Beck, C., Booth, H., El-Assady, M., & Butt, M.** (2020). Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, Barcelona, Spain (pp. 60–73). Association for Computational Linguistics (ACL). URL: <https://aclanthology.org/2020.law-1.6/>

- Berntzen, L., Johannessen, M. R., Andersen, K. N., & Crusoe, J.** (2019). Parliamentary Open Data in Scandinavia. *Computers*, 8(3), 65. DOI: <https://doi.org/10.3390/computers8030065>
- Bucks, R., Singh, S., Cuerden, J., & Wilcock, G.** (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1), 71–91. DOI: <https://doi.org/10.1080/026870300401603>
- Castiglia, T., & Walter, M.** (2008). U.S. Patent Application No. 12/041,511.
- De Clercq, O., Schulz, S., Desmet, B., Lefever, E., & Hoste, V.** (2013). Normalization of Dutch user-generated content. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2013* (pp. 179–188). Association for Computational Linguistics (ACL). URL: <https://aclanthology.org/R13-1024/>
- Erjavec, T., Grigorova, V., Ljubešić, N., Ogrodniczuk, M., Osenova, P., Pančur, A., Rudolf, M., & Simov, K.** (2020). *Multilingual comparable corpora of parliamentary debates. ParlaMint 1.0*. Slovenian language resource repository CLARIN.SI. URL: <http://hdl.handle.net/11356/1345>
- Fitsilis, F., & Koryzis, D.** (2016). 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 V*. URL: <https://www.pademia.eu/publications/online-papers-on-parliamentary-democracy/online-papers-on-parliamentary-democracy-v2016/>
- Fitsilis, F., Leventis, S., Mikros, G., & Papantonakis, P.** (2020). Xtralingua: An open-source tool for extracting quantitative text profiles. Poster presented at *Digital Humanities 2020*. DOI: <https://doi.org/10.17613/ge14-7a04>
- Fitsilis, F., Saalfeld, T., & Schwemmer, C.** (2017). Content Reconstruction of Parliamentary Questions. In *SCIECONF Proceedings*, 5(1), 107–112. DOI: <https://doi.org/10.6084/m9.figshare.14743044.v2>
- Giovani, A., Krimpas, P., Fitsilis, F., & Mikros, G.** (accepted). The terms “refugee”, “immigrant” and “illegal immigrant” in the Parliamentary Questions Corpus of the Hellenic Parliament. *To be presented in the 13th Hellenic Language and Terminology Conference* (online, 11–13 November 2021).
- Granickas, K.** (2013). *Parliamentary informatics: what data should be open and how multi stakeholder efforts can help parliaments achieve it*. European Public Sector Information Platform Topic Report No. 2013/05. URL: https://www.europeandataportal.eu/sites/default/files/report/2013_parliamentary_informatics.pdf
- Hellenic Parliament.** (2021a). *Standing Orders of the Hellenic Parliament*. URL: <https://www.hellenicparliament.gr/en/Vouli-ton-Ellinon/Kanonismos-tis-Voulis/>
- Hellenic Parliament.** (2021b). *Mésa Koinovouleftikou Elénchou [Means of Parliamentary Control]*. URL: <https://www.hellenicparliament.gr/Koinovouleftikos-Elenchos/Mesa-Koinovouleutikou-Elegxou>
- Kouklakis, G., Mikros, G., Markopoulos, G., & Koutsis, I.** (2007). Corpus Manager: A tool for multilingual corpus analysis. In D. Matthew, P. Rayson, S. Hunston & P. Danielsson (Eds.), *CL2007 Proceedings*, 27–30 July 2007, Birmingham, UK. URL: http://ucrel.lancs.ac.uk/publications/CL2007/paper/244_Paper.pdf
- Le, D. X., Straughan, S. R., & Thoma, G. R.** (2002). Greek alphabet recognition technique for biomedical documents. In *Proceedings of 6th world multiconference on systemics, cybernetics and informatics 3* (pp. 86–91). URL: <https://lhncbc.nlm.nih.gov/LHC-publications/PDF/pub2002005.pdf>
- Meinel, F.** (2018). Confidence and Control in Parliamentary Government: Parliamentary Questioning, Executive Knowledge, and the Transformation of Democratic Accountability. *The American Journal of Comparative Law*, 66(2), 317–367. DOI: <https://doi.org/10.1093/ajcl/avy028>
- Markopoulos, G., Mikros, G., Iliadi, A., & Lontos, M.** (2015). Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features. In V. Katsoni (Ed.), *Cultural Tourism in a Digital Era* (pp. 373–383). Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-15859-4_31
- Reynaert, M.** (2014). On OCR ground truths and OCR post-correction gold standards, tools and formats. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (pp. 159–166), NYC: ACM. DOI: <https://doi.org/10.1145/2595188.2595216>
- Truan, N., & Romary, L.** (2020). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. URL: <https://halshs.archives-ouvertes.fr/halshs-03097333/>
- Wang, A., Hoang, C. D. V., & Kan, M. Y.** (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1), 9–31. DOI: <https://doi.org/10.1007/s10579-012-9176-1>
- Wang, W. Y., Bohus, D., Kamar, E., & Horvitz, E.** (2012). Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Spoken Language Technology Workshop (SLT)* (pp. 73–78). Miami: IEEE. DOI: <https://doi.org/10.1109/SLT.2012.6424200>
- Zhang, J., Spirling, A., & Danescu-Niculescu-Mizil, C.** (2017). Asking too much? The rhetorical role of questions in political discourse. In *2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017* (pp. 1558–1572). Association for Computational Linguistics (ACL). DOI: <https://doi.org/10.18653/v1/D17-1164>

TO CITE THIS ARTICLE:

Fitsilis, F., & Mikros, G. (2021). Development and Validation of a Corpus of Written Parliamentary Questions in the Hellenic Parliament. *Journal of Open Humanities Data*, 7: 18, pp. 1–14. DOI: <https://doi.org/10.5334/johd.45>

Published: 04 August 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.