DiSeCan: A tool for consulting parliamentary sessions journals using linguistic intelligent text search (InteLiText)

Francisco J. Carreras-Riudavets[1,*,†] and Zenón Hernández-Figueroa[2,†]

[1] *Research Institute of Text Analysis and Applications (IATEXT), University of Las Palmas de Gran Canaria, Juan de Quesada 30, 3100 Las Palmas de Gran Canaria, Spain*

[2] *Research Institute of Text Analysis and Applications (IATEXT), University of Las Palmas de Gran Canaria, Juan de Quesada 30, 3100 Las Palmas de Gran Canaria, Spain*

Abstract

In representative democracies, parliaments hold legislative power, drafting laws and overseeing governments. Access to parliamentary proceedings empowers citizens and strengthens democracy. This document introduces DiSeCan, an online tool facilitating searches in the Session Journals of the Parliament of the Canary Islands, an autonomous community of Spain. DiSeCan offers various search methods including string-based and lexical searches, and users can filter results using metadata like speaker and session number.

Keywords

Parliamentary records, advanced search tool, linguistic annotation, information extraction

1. Introduction

In representative democracies, parliaments are the bodies entrusted with legislative power, responsible for drafting laws and overseeing the actions of governments. Effective access to knowledge about parliamentary proceedings and agreements empowers citizens and strengthens the democratic health of societies, [1] and [2] are examples of this. They are two showcases of the ParlaMint project [3], a flagship initiative by CLARIN-ERIC aimed to the creation of comparable corpora of parliamentary debates from several European countries and autonomous regions, covering at least the period from 2015 to 2022. The corpora are uniformly encoded, contain rich metadata, and are linguistically annotated up to the level of Universal Dependencies syntax and named entities.

This document presents DiSeCan [4] an online tool aimed at facilitating the search for information in the Session Journals of the Parliament of the Canary Islands, an autonomous community of the Kingdom of Spain.

The search functionality of DiSeCan supports various methods including string-based searches (exact matches, truncations, and wildcards), as well as searches based on lexical attributes such as lemma and parts of speech using a linguistic intelligent text search (InteLiText). Additionally, users can filter search results using metadata such as speaker, legislative session, session number, etc.

DiSeCan has been developed as part of the project "Ultraperiphery and European cohesion: metaphorical conceptualization of Europe in the Canarian political discourse." [5], supported by funds from the Government of the Canary Islands.

2. The Parliament of the Canary Islands

The Parliament of the Canary Islands was established with the creation of the Autonomous Community of the Canary Islands, one of the seventeen autonomous communities of Spain, through the publication of the Statute of Autonomy of the Canary Islands in August 1982 and began its first legislative session on May 30, 1983, following the elections held at the beginning of

* Corresponding author.
† These authors contributed equally.
francisco.carreras@ulpgc.es (F. J. Carreras-Riudavets);
zenon.hernandez@ulpgc.es (Z. Hernández-Figueroa);
0000-0001-9221-664X (F. J. Carreras-Riudavets); 0000-0002-1657-4020 (Z. Hernández-Figueroa);

that month. It is the legislative body of the Autonomous Community of the Canary Islands. Its main function is the drafting and approval of laws and regulations that affect the region. It is composed of deputies elected by universal suffrage in regional elections held every four years. Among its competencies are the drafting of the Statute of Autonomy, the approval of regional budgets, the oversight of the regional government, and the representation of the interests of the Canary Islands citizens.

The Session Journals of the Parliament of the Canary Islands are official records that document the proceedings, debates, decisions, and other activities that occur during the sessions of the Parliament. These journals serve as a comprehensive and accurate account of the legislative work carried out by the Parliament.

The Session Journals of the Parliament of the Canary Islands from 1983 onwards can be accessed through the parliament's own website , where they are indexed by legislature, date, and number, and can be downloaded in PDF format. Until the beginning of the third legislature in 1991, these PDFs contain TIFF images resulting from the scanning of the paper records, which require OCR processing to extract the text, with the consequent complications caused by typographical errors, distorted or slanted lines, text merging from different pages, etc. From then onwards, they contain text created directly in electronic format, which facilitates their handling. Session journals from the end of the 7th legislative term (April 2007 onwards) can also be browsed online in HTML format, a process that is intended to be extended to earlier sessions. However, in any case, the Parliament's website only allows exact text searches, justifying the development of a more powerful search tool.

3. Processing of Session Journals

The session journals are acquired in PDF format through the website of the Parliament of the Canary Islands. We have excluded session journals predating 1991 due to the low quality of the TIFF images produces a high OCR error rate. Addressing this issue would require extensive preprocessing efforts, which are impractical within the scope of the project. For session journals from 1991 onwards, which already contain electronic text, the OCR step is unnecessary as it has already been completed. The remaining OCR errors are minimal and manageable. We have also excluded the session journals from the latter part of the X Legislature due the termination of the project.

From the electronically extracted text from the PDF files, the next step is to extract the metadata that forms the general context of the session, identify the participants by assigning the corresponding text to each one, and extract the linguistic characteristics of the texts to later perform advanced textual searches. The extracted metadata and text are normalized and standardized to ensure coherence and uniformity in the dataset, and then indexed and stored in a database for quick and efficient access.

3.1. Context extraction

The metadata that forms the context of the session journal consists of the legislature number, the session number, the year and complete date, and the name of the session chairman. This information, with slight variations over time (***Figure 1*** and ***Figure 2***), is found at the beginning of the first page of the journal. Each item is easily identifiable by the presence of key expressions, such as "legislatura", "Num." or "Número," "Presidencia del Excmo. Sr. D.,", etc. which facilitates their extraction using regular expressions.

Figure 1: Context information (a).

Figure 2: Context information (b).

3.2. Speaker Identification

The time span covered by this work is very broad from a technological and methodological point of view for the Parliament of the Canary Islands. The transcripts of oral sessions have varied over the years depending on the people who performed this task. To detect the

linguistic expressions that have been used over the years to identify a speaker, several iterations of programming and manual analysis have been necessary.

Regular expressions were used to detect personal proper nouns, which must be accompanied by different expressions that indicate that the speaker is beginning to speak. The text following a personal name does not always correspond to that person as a speaker, for example, when one speaker simply mentions another. These adjacent linguistic expressions help to correctly identify the start of an intervention. The name of the political position, which the person holds in Parliament, has also been used to identify the person speaking, so that, subsequently, a conversion is made from the position to the person who holds it.

Some of the expressions considered are: "El Señor Rodríguez", "La Señora Hidalgo", "desde su escaño don Hernández", "El Señor Presidente de la Mesa", "doña Ana Marrero", etc. In the early years, "Señor" and "Señora" were mainly used before the personal proper noun, and nowadays "don" and "doña" are more common. To increase the accuracy in identifying the speaker, the personal name or position must be accompanied by the corresponding conjugated form of one of the following verbs used over the years to express the beginning of an intervention: "intervenir", "contestar", "comparecer", "replicar", "manifestar", "explicar", "hablar", "decir", "plantear", "repetir", etc., and other more complex expressions such as "hace uso de la palabra" ('take the floor').

It has also been necessary to eliminate some adverbial expressions that are found between the name and the expression: "de nuevo", "nuevamente", "impetuosamente", etc.

An additional problem that had to be solved was the different ways in which the same person can be identified, since it is customary in Spain to have two first names and two last names: by their first and first last name, by their surnames, by their first and second name and then their first last name, any part of their personal name written with or without graphic accents, any combination between the words that make up the full personal name, and even spelling mistakes and incorrect OCR results have been found. To solve this problem, all the personal proper nouns in each legislature were extracted and all the possible variants for each person were considered.

3.3. Linguistic information processing

The preprocessing of linguistic information is carried out using InteLiText (Textual Linguistic Intelligence) technology. InteLiText consists of extracting all the linguistic components of a text: paragraphs, sentences and words; and associating metadata with each text, each paragraph with its text, each sentence with its paragraph and each word with its sentence. In this extraction, the headers and footers of each page, typos, footnotes, line-ending hyphens used to break words, sentences broken on different pages, abbreviations that, despite ending in an orthographic point, do not indicate the end of a sentence, and different problems produced by the OCR process due to PDFs that were in very poor condition, mainly from the early years, have been taken into account.

To detect misspelled or misrecognized words by OCR, a Spanish lexical database (Lexicon TIP) has been used. It contains more than six million inflected (gender, number, and augmentatives) or conjugated (including enclitic pronouns) words based on more than 250,000 Spanish lemmas. Subsequently, a Part-Of-Speech Tagger (POS Tagger) is applied to all sentences, which uses the aforementioned Lexicon TIP database with detection of prefixal neologisms, to obtain the morphological information of each word in its sentence.

In summary, InteLiText technology allows the extraction of the following elements: metadata, texts, paragraphs, sentences, words, lemmas, parts of speech, and the position of words in their sentence. This information is stored in a relational database and in two text files: sentences.txt (1,522,716 sentences, 292.679KB) and paragraphs.txt (336,787 paragraphs, 257.842KB). Text files are accessed directly by position, that is, sequential access is not performed to avoid harming the performance in response speed.

The sentences.txt file contains on each line a unique numeric identifier for each sentence IDSentence, the sentence, the starting byte and the byte size of the paragraph to which the sentence belongs in the paragraphs.txt file. These last two data items allow direct access to the paragraphs.txt file. All elements are separated by a tab.

The paragraphs.txt file contains the paragraphs of all the session journals. The session journals are separated by the § character in the paragraphs.txt file, which will prevent the tool interface from displaying a context of the sentence beyond the corresponding session journal in the search results.

The structure of the sentences.txt file allows for efficient access to sentences and their corresponding paragraphs. By storing the byte offset and size of the paragraph for each sentence, the system can quickly locate the paragraph containing a given sentence without having to scan the entire paragraphs.txt file sequentially. This is especially important for large

collections of documents, where sequential access could be very slow.

The use of a tab character as the delimiter between fields in the sentences.txt file makes it easy to parse the file. Tab characters are commonly used as field separators in text files because they are not typically used in the text of the data itself. This makes it easy to split each line of the file into its individual fields using a simple string parsing routine.

Overall, the design of the sentences.txt file is well-suited for storing and accessing information about sentences in a collection of documents. The use of unique identifiers, byte offsets, and tab characters as delimiters makes the file efficient and easy to parse.

Since searches are performed based on words, lemmas, or parts of speech, the proposed design allows finding the sentences that contain those elements and the paragraphs that contain those sentences.

4. Data base

The relational database is implemented using MySQL and has a total size of 5.6GB, including indexes. It consists of three tables:

- Sessions: This table stores a unique identifier for the session (IDSession) and the metadata for each session journal. The metadata may include information such as the date of the session journal, the name of the speaker, and the topic of the discussion.
- Sentences: This table stores information about each sentence in the session journal. The information includes a unique identifier for the sentence (IDSentence), the identifier of the session journal to which the sentence belongs (IDSession), the name of the speaker who uttered the sentence, the starting byte of the sentence in the sentences.txt file, and the size of the sentence in bytes. The last two fields allow for direct access to the sentences.txt file to retrieve the full text of the sentence.
- Words: This table stores information about each individual word in the session journals. The information includes the word itself, its lemma, its parts of speech, the identifier of the sentence to which the word belongs (IDSentence), and the position of the word within the sentence.

The decision to store lemmas within the palabras table rather than creating a separate lemmas table is a trade-off between data redundancy and query performance. The design could include a table for the lemmas to avoid redundancy, because lemma is stored multiple times, once for each occurrence of the word in a sentence. However, the proposed design can improve query performance, as it eliminates the need to join two tables to retrieve a lemma for a given word. This can be especially beneficial for large datasets, where joins can be computationally expensive.

Overall, the design of the relational database appears to be well-suited for storing and retrieving information about sentences and words from session journals. The use of three tables, the storage of lemmas within the palabras table, and the use of indexes all contribute to the efficient storage and retrieval of data.

5. Tool interface

DiSeCan is a responsive web application that allows users to search for and analyse Spanish parliamentary speeches. The website is designed to be user-friendly and accessible on a variety of devices, including PCs, tablets, and smartphones. Users can search for words or phrases using a variety of contextualization criteria: date of the session journal, legislature, session number, speaker, president of the session (***Figure 3***).

Figure 3: View of DiSeCan web (smartphone).

The search results page displays a list of sentences that contain the searched-for words. For each sentence, the following information is shown:

- The paragraph where the sentence appears: This allows the user to see the sentence in its context.
- Contextualization information: This information includes the date of the session journal, the legislature, the session number, the speaker, and the president of the session. This information helps to identify and locate the sentence within the larger context of the parliamentary debate.

The size of the sentence context can be adjusted by moving the circle on the horizontal bar. Moving the circle to the left will show less context, while moving it to the right will show more context. This allows users to customize the amount of information that is displayed around the sentence.

Users can also search for phrases that contain multiple words in a specific order. By checking the "Search with order" button, users can require that the searched words appear in the sentence in the same order and with the same spacing as they were entered in the search query. This can be useful for finding exact phrases or idioms.

Overall, the search result display is designed to be informative and user-friendly. The display provides all the information that users need to understand the context of the searched sentences, and it allows users to customize the amount of context that is shown. The ability to search for phrases in a specific order is an added feature that makes DiSeCan a powerful tool for researchers and students.

Users can access the original PDF of the session journal from paragraph display. This allows users to view the full text of the speech in its original context. To access the PDF, users can click on a link that is provided next to the contextualization information (***Figure 4***).

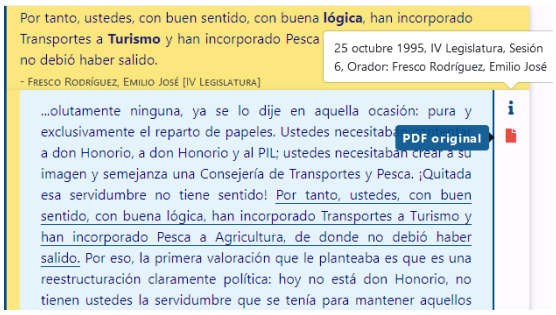

Figure 4: Search result item from DiSeCan.

This feature is particularly useful for users who want to read the speech in its entirety or who need to verify the accuracy of the quoted text. It also allows users to access additional information that may not be included in the search results, such as footnotes or appendices.

In addition to searching for exact words, you can perform the following types of advanced search:

1) Truncation search. The asterisk wildcard represents any sequence of characters. For example, the search query "universi*" would return results that include the words "universidad", "universidades", "universitario", "universitarios", and so on.

2) Lemma search. Searches for all the inflected forms of a verb (with or without enclitics) or the gender and number inflections of any word (including appreciative forms: augmentatives, diminutives, and pejoratives), even if they have prefixes: “pretratar” (“tratar” verb with prefix *pre-* and enclitic pronoun *-las*). The $ character is used at the beginning of the word to indicate this type of search. If the word is not a lemma, its lemma is automatically determined before the search is performed. For example, *$dímelo* will be automatically changed to *$decir* (“decir” is lemma from “dímelo”) before performing the search in the corpus.

3) Part of speech search. Searches for phrases that contain words with a specific part of speech: <part of speech> is written between angle brackets. The accepted parts of speech are adjective, adverb, article, conjunction, interjection, preposition, pronoun, noun, and verb.

4) Proximity search. Searches for words within a specified distance of each other in phrases. This type of search considers the order of the words in the phrase. Proximity search, also known as nearness search, allows users to find words that are close to each other in a phrase. This can be useful for a variety of purposes, such as: identifying collocations, analysing sentence structure, and extracting information from text. You can specify the maximum distance between the target words with a number.

Likewise, all search options can be combined according to the user's needs. For example, "$te 3 miedo <preposición>" searches for all inflections or conjugations of any lemma that starts with "te" and has a maximum of three words before finding "miedo" followed by a preposition. The result of the previous search will be phrases that contain expressions of the type:

- *teniendo miedo en*
- *tuviera miedo a*
- *tienen menos miedo de*

- *tendría tanto miedo a*
- *término sin ningún miedo en*
- *terrorismo y el miedo a*

To demonstrate the potential and efficiency of DiSeCan, consider the following example queries that showcase its ability to handle complex information retrieval tasks:

• *What was discussed in the 8th Legislature about the Government of Spain?*

You must select VIII legislatura and write: *Gobierno 2 España*. Number 2 allows you to find sentences with the following sequences. "*Gobierno Socialista de España*", "*Gobierno democrático de España*", "*Gobierno legítimo de España*", "*Gobierno que cohesione España*", among other possibilities.

• *What did Paulino Rivero say about "trabajo"?*

You must select Paulino Rivero speaker and write the lemma preceded by the dollar sign: *$trabajo*. Find the sentences that contain any inflection of "trabajo" ("trabajo", its plural, and any appreciative forms: "trabajillo", "trabajitos", etc.: *Porque, claro, cuando para el empresario todo son riesgos, todo son dificultades, todo son obstáculos, pues dice "yo, más cómodo, me empleo y si consigo un trabajito tranquilo, mejor"*.

• *$rey 2 Carlos*

Find the sentences that contain any inflection of "rey" ("rey", "reyes", "reina", "reinas", "reyezuelo", reinitas, etc), followed by a maximum of two words before "Carlos" word. Number 2 allows you to find different options ("rey Juan Carlos" and "rey don Juan Carlos"): *"Por ejemplo, fue una expresión de refuerzo de la monarquía española la abdicación del rey Juan Carlos a favor del Rey Felipe VI."*

• *$decir 5 Gobierno*

Find the sentences that contain any verbal form of "*decir*" (include enclitic pronouns), followed by a maximum of five words before "Gobierno" word: *"Permítame que le diga señor Presidente del Gobierno que es más posible llegar con su política a Cánovas que al preámbulo de la Constitución cuando habla de una democracia avanzada."*, *"He venido a reflexionar a tenor de lo que aquí se ha dicho de los diez años de gobierno de la coyuntura económica de la situación social que el cambio sigue siendo necesario."*

• *Gobierno <adjetivo>*

Finde the "*Gobierno*" word followed by an adjetive: *"El terrorismo considera al Gobierno democrático como un Gobierno enemigo y, por consiguiente, golpea al Gobierno y a la sociedad, simplemente porque no quiere que se viva en libertad."*, *"Desde luego quiero afirmar que este Gobierno es el Gobierno legítimo de España y que contra él no se puede hacer nada fuera de la Constitución."*

6. Future work

6.1. Metadata expansion

Currently, for the purposes of the project, we have extracted a limited set of metadata fields from the session journals of the Canary Islands Parliament, but there are more that could be gathered and added as filters or search criteria.

For example, concerning the speakers, data such as gender, age (date of birth), political group, etc., could be provided, which could be useful as variables to correlate in the analysis of their discourse. The reason for not having considered these kinds of data in the current state of the project is that, unlike those that have been used, they cannot be directly extracted from the text of the session journals so requiring other sources.

6.2. Semantic relations

The TIP Lexicon classifies over 80,000 derivational relationships that InteLiText can use in the linguistic preprocessing to enable searches for word families. For example, when the user types the word "proteger" (to protect), he can obtain results with sentences containing the words "proteger", "protección" (protection), "protector", "protegido" (protected), "autoproteger" (self-protect), "desproteger" (unprotect), among others.

InteLiText can extend the type of semantic relationships to synonyms, antonyms, hypernyms, hyponyms, etc. based on the relationships established in WordNet for Spanish [7].

6.3. Automation of corpus collecting

In the context of a new project, incorporating the most recent documents omitted after the completion of the previous one into the DiSeCAn corpus is relatively straightforward. However, including documents from the initial legislatures, which necessitate greater effort due to their low quality, is a bit more challenging but still feasible. As session diaries will continue to accumulate, it would be advisable to develop a tool capable of automatically processing new documents as they are published on the Canary Islands Parliament website. This is feasible since the higher quality of the new documents will require, at most, minimal human supervision.

6.4. Exporting machine readable corpora

As demonstrated by the ParlaMint project, open-access datasets unlock a myriad of research opportunities and facilitate interdisciplinary studies. Therefore, while DiseCan enables powerful searches, providing scholars with access to machine-readable corpora could further enhance their research capabilities.

The Parla-CLARIN recommendations [8] for encoding parliamentary corpora compiled for scholarly research could be useful for this task. The Parla-CLARIN recommendations were developed at an initial stage of the ParlaMint project whose schema is compatible with them.

6.5. Importing machine readable corpora

Being able to implement a scheme based on the Parla-CLARIN recommendations for exporting the DiSeCan corpus, it would be feasible to develop the import of other parliamentary corpora following similar schemes. For instance, corpora produced by projects like ParlaMint could be imported, enabling the creation of a tool inspired by DiSeCan. This tool could search multiple corpora simultaneously or separately.

References

[1] ParlaMint and ParlaMeter: How Standardised Data Formats Empower End Users. Filip Dobranić, CLARIN Café: ParlaMint Unleashed, 2021.

[2] ParlaMint - A Resource for Democracy. Dario Del Fante and Virginia Zorzi, 'Who Is the Enemy Now?', CLARIN Impact Stories, 2023.

[3] ParlaMint: Comparable and Interoperable Parliamentary Corpora. URL: https://www.clarin.eu/parlamint.

[4] DiSeCan - Buscador-IL en el Diario de Sesiones del Parlamento de Canarias. URL: https://dise.iatext.ulpgc.es/canarias/.

[5] Ultraperiferia y cohesión europea: conceptualización metafórica de Europa en el discurso político canario. URL: https://iatext.ulpgc.es/es/ultraperiferia_cohesion_europea.

[6] Parlamento de Canarias. URL: https://www.parcan.es/.

[7] Fernández-Montraveta, Ana & Vázquez, Gloria & Fellbaum, Christiane. (2008). The Spanish Version of WordNet 3.0. doi: 10.1515/9783110211818.3.175.

[8] Erjavec, T., & Pančur, A. The Parla-CLARIN Recommendations for encoding corpora of parliamentary proceedings. Journal of the Text Encoding Initiative 14 (2021). doi: 10.4000/jtei.413.