Digitization of the Australian Parliamentary Debates (1901-2022)

Lindsay Katz Rohan Alexander

October 17, 2022

Parliamentary dialogue is critical for political science research. In Australia, following the UK tradition, the written record of what is said in parliament is known as Hansard. While the Australian Hansard has always been publicly available, it has been difficult to use it for the purpose of large-scale macro and micro-level text analysis because it has not been available as a dataset of sufficient quality to be credibly analysed with statistical models. Following the lead of the Lipad project which achieved this for Canada, our project aims to provide a new, comprehensive, high-quality database that captures all proceedings of the Australian parliamentary debates from 1901 to the present day using Hansard. Our dataset will be publicly available and linked to other datasets such as election results. The creation of this dataset will enable the exploration of questions that are not currently possible to explore, serving as a valuable resource for both researchers and policymakers. This work will provide a thorough description of the creation and computational underpinnings of this database, followed by a discussion of some applications.

1 Introduction

—— need better intro bit here —— The official written record of parliamentary debates, formally known as Hansard, play a fundamental role in capturing the history of political proceedings and facilitating the exploration of valuable research questions. Originating in the British parliament, the production of these reports became tradition in a number of Commonwealth countries such as Canada, the United Kingdom, and Australia (Vice and Farrell 2017). Given the contents and sheer magnitude of these records, their value cannot be overstated - particularly in the context of political science research. In the case of Canada, a team of researchers at the University of Toronto have digitized Hansard from 1901 to 2019, an endeavor called the Linked Parliamentary Data (LiPaD) project (Beelen et al. 2017). Having a digitized version of Hansard enables researchers to perform advanced analyses on these records using text analysis tools and statistical modelling. Inspired by the LiPaD project, in this paper we introduce a novel database for Australian Hansard. This is composed of individual datasets for each sitting day in the House of Representatives, containing details on everything said in parliament in a form which can be readily used by researchers. With the development of tools for large-scale text analysis, this database will serve has a valuable resource for studying and exploring political behaviour in Australia over time.

The House of Representatives performs a number of crucial governmental functions, such as creating new laws and overseeing government expenditure (House of Representatives, Elder, and Fowler 2018, ch. 1). Sittings of the House in the follow a predefined order of business, regulated by procedural rules called standing orders (House of Representatives, Elder, and Fowler 2018, ch. 8). A typical sitting day in the Chamber includes debates on government business, 90 second member statements, Question Time, document presentation, matters of public importance, ministerial statements, and an adjournment debate (House of Representatives, Elder, and Fowler 2018, ch. 8). In contrast to the Chamber, sittings in the Federation Chamber are quite different in terms of their order of business and scope of discussion. The Federation Chamber was created in 1994 as a subordinate debate venue, to allow for better time management of House business as its proceedings occur simultaneously with those of the Chamber (House of Representatives, Elder, and Fowler 2018, ch. 21). Business matters discussed in the Federation Chamber are limited largely to intermediate stages of bills, and the business of private Members (House of Representatives, Elder, and Fowler 2018, ch. 21). It is the compilation of these proceedings upon which Hansard is based.

On each sitting day, a transcript is available for download from the official Parliament of Australia website in both scanned PDF and extensible markup language (XML) form. The scanned PDF is the official release, which is converted to typed text using Optical Character Recognition (OCR) technology (Sherratt 2016). This conversion allows for the production of the XML formatted transcript, though these are not always perfect conversions. Our database has been developed solely using the XML formatted files. The choice to create this database using the potentially incomplete or flawed XML format of Hansard as opposed to the scanned PDF is —- mention previous work on this using the PDF and this motivated us to try to parse things from the XML side this time —— .

cases where unsure how to proceed w xml always deferred to pdf

The end goal for this work is to create a publicly available database of parsed Hansard transcripts from 1901 to present in a form which can be readily analyzed. More specifically, each transcript is presented in an ordered form with details on each speaker such as their political affiliation, the time stamp of their statement, and whether they interjected (i.e. spoke out of turn). Thus far, this has been completed from 2011 to present. Due to the extensive variation in structural formatting of these transcripts, additional work is needed to parse transcripts from further in the past.

We begin this paper with a description of what datasets for Australian Hansard already exist will be provided (Section 2). Next will be a detailed description of the contents of our database

(Section 3), followed by a thorough discussion of how this database was created (Section 4). Finally, Section 5 provides a description of some applications of our database.

2 Existing Datasets

- What already exists, and how has it been used
- How are these existing data sets lacking?
- How do we account for this in our data set?
    - WhoGov article has good summary table offering comparison to other existing data sets

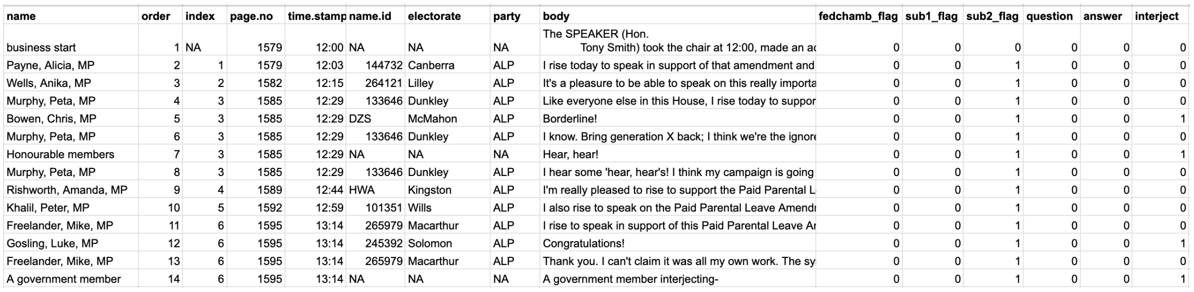
to this point various researchers have had to create

- tim sherrett (historic hansard) -
- patrick leslie has also created a dataset as have others rework paragrpah about how these have been used

no comprehensive dataset based on xml that goes from 1901-today 1930s and before - ocr ruins accuracy (earlier decades)

3 Description of the Database

Our database thus far contains 1388 CSV files, one for each sitting day of the House of Representatives from 10 May 2011 to 08 September 2022. Each CSV is produced from the same script, so all files in our database are formatted and processed in the exact same way. Figure 1 provides a snapshot of one of these files.



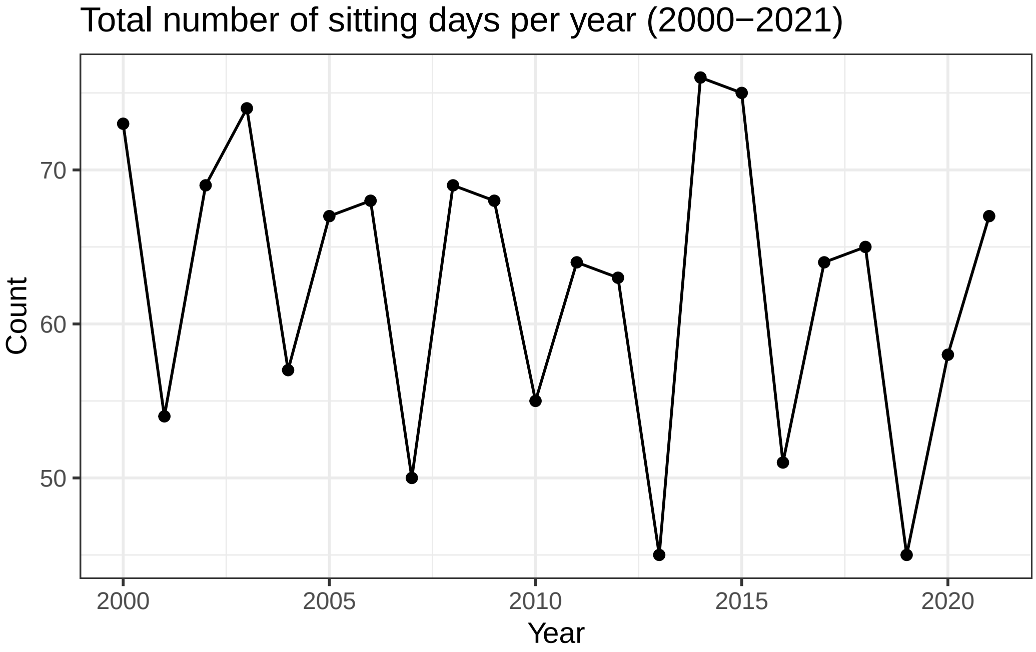
name	order	index	page.no	time.stamp	name.id	electorate	party	body	fedchamb_flag	sub1_flag	sub2_flag	question	answer	interject
business start	1	NA	1579	12:00	NA	NA	NA	The SPEAKER (Hon. Tony Smith) took the chair at 12:00, made an ac	0	0	0	0	0	0
Payne, Alicia, MP	2	1	1579	12:03	144732	Canberra	ALP	I rise today to speak in support of that amendment and	0	0	1	0	0	0
Wells, Anika, MP	3	2	1582	12:15	264121	Lilley	ALP	It's a pleasure to be able to speak on this really importa	0	0	1	0	0	0
Murphy, Peta, MP	4	3	1585	12:29	133646	Dunkley	ALP	Like everyone else in this House, I rise today to suppor	0	0	1	0	0	0
Bowen, Chris, MP	5	3	1585	12:29	DZS	McMahon	ALP	Borderline!	0	0	1	0	0	1
Murphy, Peta, MP	6	3	1585	12:29	133646	Dunkley	ALP	I know. Bring generation X back; I think we're the ignore	0	0	1	0	0	0
Honourable members	7	3	1585	12:29	NA	NA	NA	Hear, hear!	0	0	1	0	0	1
Murphy, Peta, MP	8	3	1585	12:29	133646	Dunkley	ALP	I hear some 'hear, hear's! I think my campaign is going	0	0	1	0	0	0
Rishworth, Amanda, MP	9	4	1589	12:44	HWA	Kingston	ALP	I'm really pleased to rise to support the Paid Parental L	0	0	1	0	0	0
Khalil, Peter, MP	10	5	1592	12:59	101351	Wills	ALP	I also rise to speak on the Paid Parental Leave Amendm	0	0	1	0	0	0
Freelander, Mike, MP	11	6	1595	13:14	265979	Macarthur	ALP	I rise to speak in support of this Paid Parental Leave Am	0	0	1	0	0	0
Gosling, Luke, MP	12	6	1595	13:14	245392	Solomon	ALP	Congratulations!	0	0	1	0	0	1
Freelander, Mike, MP	13	6	1595	13:14	265979	Macarthur	ALP	Thank you. I can't claim it was all my own work. The sy	0	0	1	0	0	0
A government member	14	6	1595	13:14	NA	NA	NA	A government member interjecting-	0	0	1	0	0	1

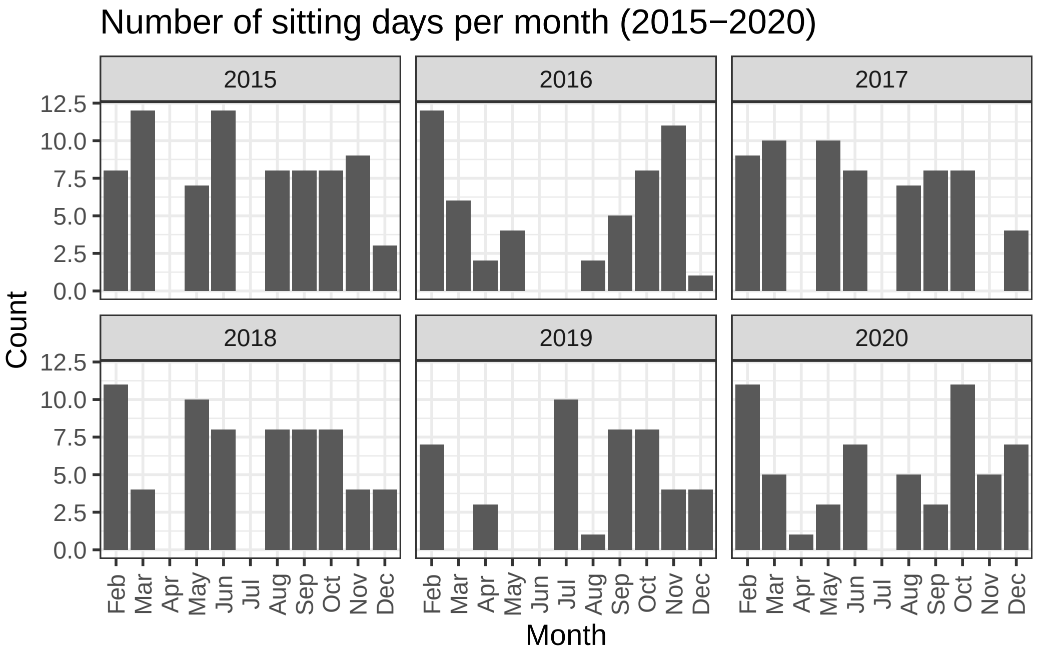
Figure 1: First 15 rows of CSV file for Hansard on 25 February 2020

Variable	Description
name	Name of speaker

Variable	Description
order	Row number
index	Speech number, to keep track of which separated text belonged to which original speech
page.no	Page number statement can be found on in official Hansard
time.stamp	Time of statement
name.id	Unique member identification code
electorate	Speaking member's electorate
party	Speaking member's party
body	Statement text
fedchamb_flag	Flag for Federation Chamber (1 if Federation Chamber, 0 if Chamber)
sub1_flag	Flag for sub-debate 1 contents (1 if sub-debate 1, 0 otherwise)
sub2_flag	Flag for sub-debate 2 contents (1 if sub-debate 2, 0 otherwise)
question	Flag for question (1 if question, 0 otherwise)
answer	Flag for answer (1 if answer, 0 otherwise)
interject	Flag for interjection (1 if statement is an interjection, 0 otherwise)

- number missing (so far two days in 2014)
- number unique speakers
- avg number of rows per day
- number of interjections (graph over time?)





4 Creation of the Database

The approach to parsing contents of an XML document is heavily dependent on its tree structure. As such, to create this database, we started by looking at a single Hansard transcript in XML form. Doing so enabled us to understand the various components of interest in the document, and how each one can be parsed according to its corresponding structural form. For example, the beginning of a speech may be followed by various interjections and continuations, all of which are structured as unique sub-elements containing further nested sub-elements within them. Parsing was performed in `R` using the `XML` package. Focusing on one transcript also allowed us to ensure that all key components of the transcript were being parsed and captured with as much detail as possible. Further, since we are looking at such a wide time span of documents, there are expectedly a number of changes in the way they are formatted. Some changes are as subtle as a differently named child-node, while others are as extensive as a completely different nesting structure. Smaller changes were accounted for as we became aware of them, and embedded into the final scripts in a way that would not cause issues for parsing more current Hansards with slightly different formatting. As mentioned in Section 1, the significant changes in XML structure of Hansard preceding 2011 necessitates much further development of our current scripts, which we are actively working to complete.

—— should I add little example of the XML to give the reader a visual?? it's hard to put into words clearly ——

The XML structure begins with a root or parent element `"hansard"`, followed by a child element `"session.header"` with sub-child elements with information such as the date, and parliament number. Next there is a child element containing everything that takes place in the Chamber, `"chamber.xscript"`. On sitting days where the Federation Chamber met as well, a child element exists for the Federation Chamber too. This allows for the distinction of proceedings of the Chamber and Federation Chamber. Each proceeding begins with a business start, naming the speaker and what time they took the chair. Next comes the debates and nested sub-debates. Broadly, the speeches within Hansard transcripts are structured with a debate node, containing a sub-debate 1 child-node which has a sub-debate 2 child-node nested within it. That said, sometimes sub-debate 2 is not nested within sub-debate 1. Each of these three elements as well as their respective sub-elements contain important information on the title of the debate or sub-debate, who is speaking, and what is being said. As mentioned, in many cases speeches are interrupted by interjections, which are embedded in these sub-elements as well. The final key distinct component of Hansard is Question Time, in which question and answer elements are classified differently than general debate text. Sometimes questions and answers are nested within debate or sub-debate child nodes, and other times they have their own child node. More detail on the processing of Question Time will follow in Section 4.1.

Once code was written to parse all components, it was organized into individual scripts. The first script includes everything from the session header element, the second contains everything from the debate element, and the third and fourth contain everything from the sub-debate 1 and 2 elements, respectively. The fifth script contains debate interjection information, and the sixth contains content from Question Time. The final script is a compilation of everything. The next step was to further develop these scripts to produce tidy data sets from each parsed element, where each statement is separated onto its own row with details about the speaker, and rows are placed in chronological order. This first involved correcting the variable classes and adding a number of indicator variables to differentiate where statements came from, such as Chamber versus Federation Chamber or sub-debate 1 versus sub-debate 2. The next key task revolved around the fact that the raw text data were not separated by each statement when parsed. In other words, any interjections, comments made by the Speaker and Deputy Speaker and continuations within an individual speech were all parsed together as a single string. As such, the name, name ID, electorate and party details were only provided for the person who's turn it was to speak. Splitting up these speeches in a way which would be generalizable across sitting days required much thought and effort. Section 4.2 will provide further details on the intricacies of this task.

When running these scripts on transcripts from other sitting days, it became clear that not every sitting day contains every possible XML element. For example, some days did not have sub-debate 2 content, and some days did not have a Federation Chamber proceeding. To improve the generalizability of these scripts, if-else statements were embedded within the code wherever an error might arise due to a missing element. For example, the entire Federation Chamber block of code is wrapped in an if-else statement for each script, so that it only executes if what the code attempts to parse actually exists in the file. Once the script ran without error for a few recent years of Hansard, we continued to work backwards until extensive changes in tree structure made our script incompatible with parsing earlier XML files. Specifically, the earliest date for which it works is 10 May 2011. Before writing a new script for parsing earlier Hansard, we decided to prioritize cleaning and finalizing what we have been able to parse. As such we continued building our scripts, fixing any issues we noticed in the resulting datasets, and separating any additional sections of the parsed text where necessary. Specifically, we added a section of our script to separate on general stage directions. More information on this separation will be provided in Section 4.3.

4.1 Question Time

A key characteristic of the Australian parliament system is the ability for the executive government to be held accountable for their decisions. One core mechanism by which this is achieved is called Question Time. This is a period on each sitting day in the Chamber where members of the House can ask ministers two types of questions: questions with notice (often referred to as questions in writing), which are written in advance, or questions without notice, which are asked verbally in the Chamber and are responded to in real time (Representatives 2021). Parsing the components of Question Time required a slightly different approach than that of the debate speech, because of its unique structure in the XML document. Sometimes, questions in writing are included directly in the "chamber.xscript" child node, with sub-child nodes called "question" and "answer" to differentiate the two. However, in other times, questions in writing are embedded in their own child node called "answers.to.questions" outside of "chamber.xscript". Difficulty arose when trying to order these questions and answers correctly within the rest of the debate text, because unlike all other statements made which have both an associated page number and time stamp, these generally only have associated page numbers. Time stamps have been a key variable for which we have arranged parsed text in chronological order, especially since many statements tend to be on the same page number, meaning we cannot rely solely on page number to capture correct ordering. To approach this issue, we merged all parsed questions to be in a single dataframe, and all parsed answers in another dataframe. We then arranged each dataframe by page number, and merged the two dataframes in such a way that each question would be followed by it's corresponding answer.

4.2 Interjections

- First thought was to use a number of phrases, but it would be really difficult to create an exhaustive list manually
- Once all the text were split to have a different row for each interjection/continuation etc, we could just flag for interjections automatically b/c i kept the original row number that the speech started on so I could flag if the original speaker or speaker/deputy speaker wasnt speaking, it was an interjection

- Had to consider a ton of different names and possible formatting of the names
  - Ex: last name only, two first names, last names with two words or hyphen or apostrophe, people with the same last name
  - had to also find a way to match these various formats to the name id, electorate, party info
- Also had to consider general interjections that are not directly named, such as “an opposition member” or “government members”, or “Mr Smith interjecting-”
- Some people have no first name so couldn’t manually extract their info (i.e. electorate, party etc.)

Tons of variation in how names are formatted (Capitalization makes it super difficult with regex) - also pre 2021 the speaker is not fully identified the same way as in 2021 2022 - just lots of small changes to account for

4.3 Stage Directions

As mentioned, one of the final components added to our script was to separate general stage directions out from statements made by members. Stage directions are general statements included in the transcript to document happenings in the Chamber. Examples of stage directions are “Bill read a second time”, “Question agreed to” or “Debate adjourned”. It is unclear from the XML and PDF who exactly these statements are attributed to, or in other words, who are making these statements. For further clarification, we watched portions of the video recording for some sitting days, and noticed that where these statements are documented in Hansard, they are not explicitly stated in parliament. For example, when the Deputy Speaker says “The question is that the bill be now read a second time”, members of the House take a vote, and if the majority is in favour, they proceed reading the bill the second time - but this is not explicitly transcribed, rather what is written is: “Question agreed to. Bill read a second time”. For this reason, we filled the name variable for these statements with “stage direction”. It is important to note that these stage directions are not defined differently from the regular debate speech in the XML, meaning we had to manually build a list of stage directions to separate out. We have been building this list of stage directions as we work backwards in parsing Hansard.

5 Applications

6 Conclusion

References

Beelen, Kaspar, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, et al. 2017. “Digitization of the Canadian Parliamentary Debates.” *Canadian Journal of Political Science/Revue Canadienne de Science Politique* 50 (3): 849–64.

House of Representatives, Department of the, D. R. Elder, and P. E. Fowler. 2018. *House of Representatives Practice: 7th Edition*. Australian Government - Department of the House of Representatives. https://books.google.ca/books?id=XzYgtAEACAAJ.

Representatives, House of. 2021. *A Window on the House: Practices and Procedures Relating to Question Time*. Parliament of Australia.

Sherratt, Tim. 2016. “Documentation: Historic Hansard.” *Historic Hansard*. http://timsherratt.org/digital-heritage-handbook/docs/historic-hansard/#.

Vice, John, and Stephen Farrell. 2017. *The History of Hansard*. House of Lords Library; House of Lords Hansard.