Digitization of the Australian Parliamentary Debates (1901-2022)

Lindsay Katz **Rohan Alexander**

University of Toronto

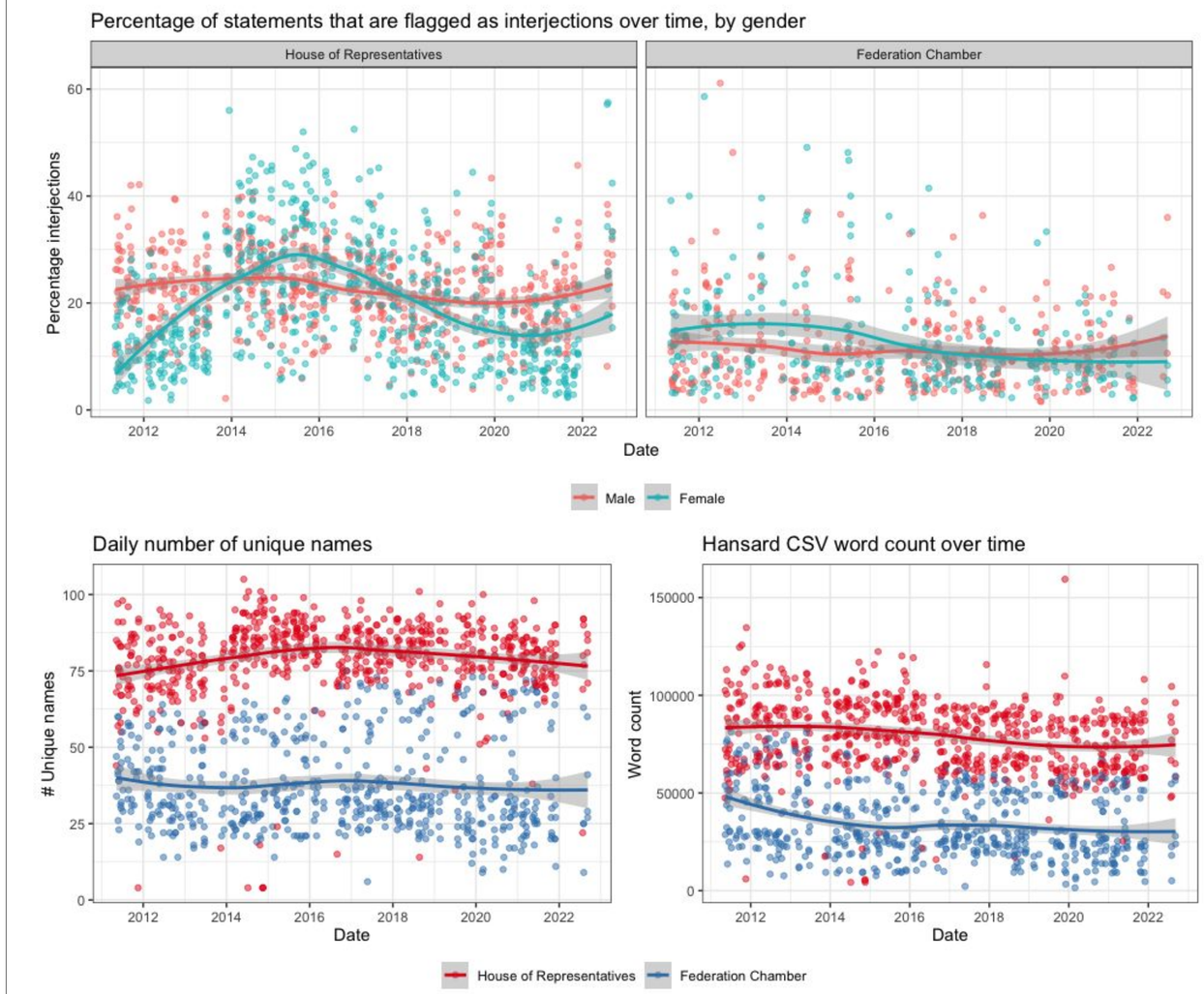
Overview

In our work, we introduce a novel database for Australian Hansard, which captures all proceedings of the House of Representatives from 1901-2022. Our database currently contains 669 CSV files, one for each sitting day of the House of Representatives from 10 May 2011 to 08 September 2022.

Method

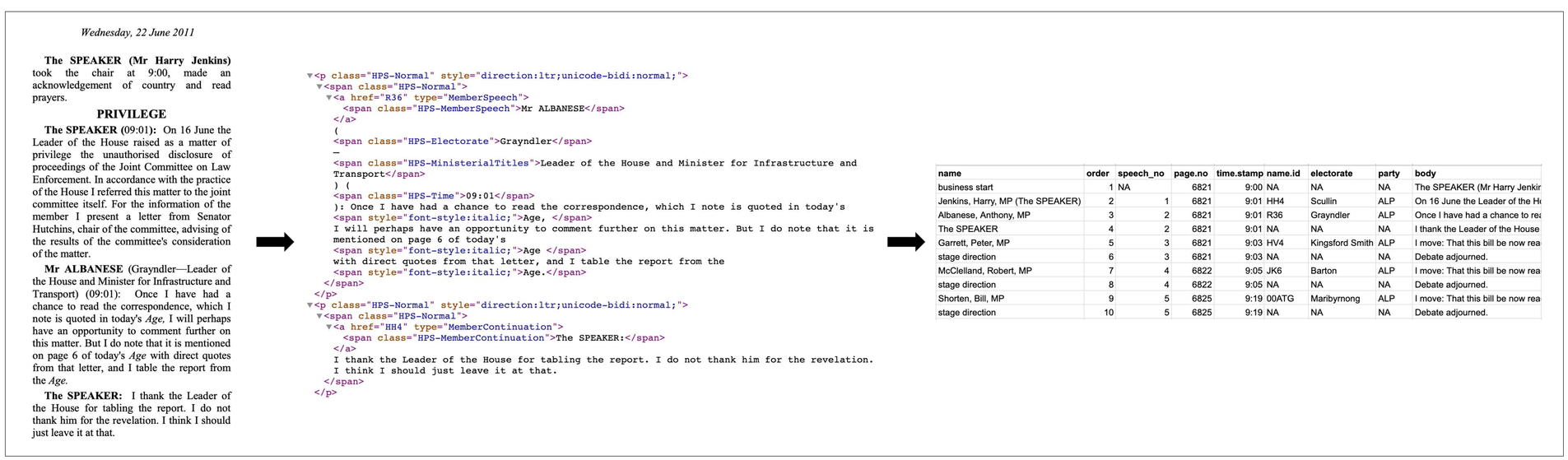
1. Work through a single XML to identify components of interest and their corresponding structural form
2. Write code to parse each component
3. Develop script to combine all components chronologically, and separate all statements onto their own rows
4. Generalise script to parse other sitting days, working backwards in time and adding to script as needed

Variable	Description
name	Name of speaker
order	Row number
speech_no	Speech number
page.no	Page number from official Hansard PDF
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	1 if Federation Chamber, 0 if Chamber
sub1_flag	1 if sub-debate 1 content, 0 otherwise
sub2_flag	1 if sub-debate 2 content, 0 otherwise
question	1 if question, 0 otherwise
answer	1 if answer, 0 otherwise
q_in_writing	1 if question in writing, 0 otherwise
interject	1 if statement is an interjection, 0 otherwise



Main Challenges

- Raw text data were not separated by statement when parsed. So, any interjections, continuations, and comments within an individual speech were parsed as a single string, for example:
  - "Mr ANDREWS (Menzies) (10:38): These are—Mr Martin Ferguson: Talk about flat-footed. Mr ANDREWS: I was just interested in the interjections by the Minister for Resources and Energy. I am always interested in anything that the minister, who is across the table, says.The DEPUTY SPEAKER (Mr KJ Thomson): The member should not be interested in interjections.Mr Martin Ferguson interjecting—"
- This required us to consider all possible forms in which names and interjections could be transcribed, to separate the text in the correct spots
  - Examples of some variations in names:
    - Mr McCormack
    - Mr Michael McCormack
    - Mr Michael McCORMACK
    - McCormack, Michael, MP
    - Mr McCormack interjecting-
    - Mr Michael McCormack interjecting-
- Sometimes, Members were simply being named in the statement of another Member. Regular expression lookarounds allowed us to minimize incorrect separations of text such as these
- We then had to match names with the correct full name, name ID, electorate, and party data
- We generalized our code to account for changes in XML formatting and transcription style over time
  - Changes included the style of timestamp transcription, name transcription, and the names of XML elements



Wednesday, 22 June 2011

The SPEAKER (Mr Harry Jenkins) took the chair at 9:00, made an acknowledgement of country and read prayers.

PRIVILEGE

The SPEAKER (09:01**):** On 16 June the Leader of the House raised as a matter of privilege the unauthorised disclosure of proceedings of the Joint Committee on Law Enforcement. In accordance with the practice of the House I referred this matter to the joint committee itself. For the information of the member I present a letter from Senator Hutchins, chair of the committee, advising of the results of the committee's consideration of the matter.

Mr ALBANESE (Grayndler—Leader of the House and Minister for Infrastructure and Transport) (09:01): Once I have had a chance to read the correspondence, which I note is quoted in today's *Age*, I will perhaps have an opportunity to comment further on this matter. But I do note that it is mentioned on page 6 of today's *Age* with direct quotes from that letter, and I table the report from the *Age*.

The SPEAKER: I thank the Leader of the House for tabling the report. I do not thank him for the revelation. I think I should just leave it at that.

```
<p class="HPS-Normal" style="direction:ltr;unicode-bidi:normal;">
 <span class="HPS-Normal">
  <a href="R36" type="MemberSpeech">
   <span class="HPS-MemberSpeech">Mr ALBANESE</span>
  </a>
  (
  <span class="HPS-Electorate">Grayndler</span>
  —
  <span class="HPS-MinisterialTitles">Leader of the House and Minister for Infrastructure and
  Transport</span>
  ) (
  <span class="HPS-Time">09:01</span>
  ): Once I have had a chance to read the correspondence, which I note is quoted in today's
  <span style="font-style:italic;">Age, </span>
  I will perhaps have an opportunity to comment further on this matter. But I do note that it is
  mentioned on page 6 of today's
  <span style="font-style:italic;">Age </span>
  with direct quotes from that letter, and I table the report from the
  <span style="font-style:italic;">Age.</span>
 </span>
</p>
<p class="HPS-Normal" style="direction:ltr;unicode-bidi:normal;">
 <span class="HPS-Normal">
  <a href="HH4" type="MemberContinuation">
   <span class="HPS-MemberContinuation">The SPEAKER:</span>
  </a>
  I thank the Leader of the House for tabling the report. I do not thank him for the revelation.
  I think I should just leave it at that.
 </span>
</p>
```

name	order	speech_no	page.no	time.stamp	name.id	electorate	party	body
business start	1	NA	6821	9:00	NA	NA	NA	The SPEAKER (Mr Harry Jenkir
Jenkins, Harry, MP (The SPEAKER)	2	1	6821	9:01	HH4	Scullin	ALP	On 16 June the Leader of the H
Albanese, Anthony, MP	3	2	6821	9:01	R36	Grayndler	ALP	Once I have had a chance to rea
The SPEAKER	4	2	6821	9:01	NA	NA	NA	I thank the Leader of the House
Garrett, Peter, MP	5	3	6821	9:03	HV4	Kingsford Smith	ALP	I move: That this bill be now rea
stage direction	6	3	6821	9:03	NA	NA	NA	Debate adjourned.
McClelland, Robert, MP	7	4	6822	9:05	JK6	Barton	ALP	I move: That this bill be now rea
stage direction	8	4	6822	9:05	NA	NA	NA	Debate adjourned.
Shorten, Bill, MP	9	5	6825	9:19	00ATG	Maribyrnong	ALP	I move: That this bill be now rea
stage direction	10	5	6825	9:19	NA	NA	NA	Debate adjourned.