Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 1.1 (2019) 1–43
HTTPS://DOI.ORG/10.5117/CCRXXXXXXXXX

Hansard DB: A Relational Database of Australian Parliamentary Speech

Alfie Chadwick[1], Simon D. Angus[2], Libby Lester[3]

[1] *Monash Climate Communication Hub, School of Media, Film and Journalism, Monash University*

[2] *Dept. of Economics & SoDa Laboratories, Monash Business School, Monash University*

[3] *Monash Climate Communication Hub, School of Media, Film and Journalism, Monash University*

Abstract

Legislative chambers are central institutions of democratic governance, where representatives debate policy, justify decisions, and are held accountable. In Australia, the principal record of what is said and done in parliament is the official parliamentary transcript, commonly known in Westminster systems as Hansard. Yet while these records are publicly accessible, they are difficult to use at scale because they are not readily queryable, computationally integrated, or suitable for systematic longitudinal analysis.

This paper addresses that problem by introducing Hansard DB, a relational database of Australian federal parliamentary speech spanning Federation to the present. Hansard DB integrates speeches, questions, answers, and interjections with parliamentarian metadata from the Parliamentary Handbook. Built through a multi-stage parsing and validation pipeline, the database supports flexible querying across text and metadata for large-scale and longitudinal analysis.

The paper also discusses the epistemic limitations of Hansard research. The official transcripts are edited and incomplete records of political speech, shaped by institutional rules, editorial practices, and omissions of tone, gesture, and context. Hansard DB therefore contributes not only a usable database, but also guidance for interpreting parliamentary transcripts acconting for these limitations.

Keywords: To Insert

Introduction

In any modern democracy, transparency is essential to ensuring accountability, public trust, informed participation, and the prevention of corruption (Stiglitz, 1999). One of the primary instruments for transparency in Australia – and in many other Westminster-style democracies – is Hansard: the official, verbatim transcript of parliamentary proceedings (Parliament of Australia, n.d.-a). Hansard provides a detailed public record of every division, question, and speech delivered in both houses of the Australian Parliament since Federation in 1901 and is an invaluable resource for citizens, journalists, and historians to scrutinise debates and hold elected representatives accountable.

However, the current form of Hansard, while publicly available, presents significant challenges for those wishing to conduct large-scale analysis. Parliamentary records are published in PDF or XML formats, both of which often suffer from inconsistent formatting and structural errors. These issues make automated data extraction and processing difficult, particularly for researchers and developers seeking to build tools or undertake comprehensive analyses. Furthermore, Hansard's official repository, ParlInfo, offers limited functionality: it does not support complex querying of the full corpus or easy extraction of targeted subsets of data.

The challenges surrounding the accessibility of the Australian Hansard have led to a highly fragmented research environment for Australian parliamentary texts. Researchers often find themselves creating bespoke datasets – each extracting and cleaning Australian Hansard data according to their own requirements and unique methodological approaches (Alexander & Alexander, 2021; Hames et al., 2025; Martínez Arranz et al., 2023; Ng et al., 2025). This repeated, individualised extraction not only leads to unnecessary duplication of effort, but also results in a proliferation of datasets that are difficult to compare or integrate across projects.

This problem is particularly acute given the extraordinary advances in computational tools for textual analysis over the past decade. Natural language processing (NLP), including transformer-based models and large language models (LLMs), have opened new possibilities for extracting meaning, sentiment, argument structure, and rhetorical patterns from large corpora (Ziems et al., 2024). Sophisticated discourse analysis platforms have become increasingly accessible, enabling researchers across communications, political science, economics, and digital humanities to pursue questions that were

previously impractical or impossible to investigate at scale. Yet the utility of these powerful analytical tools depends fundamentally on the availability of data in structured, machine-readable formats. None of these capabilities matter if the underlying data cannot be reliably extracted, cleaned, and organised for analysis (Angus, 2026).

Katz and Alexander (2023) made an important contribution toward addressing these challenges by providing a tabular dataset of Australian Hansard proceedings. This resource significantly improved longitudinal research by offering cleaner and more reliable data than raw downloads. But important limitations remain. Current coverage begins in 1998 (Katz et al., 2026), and rectangular data formats are less well suited to representing the nested and relational structure of parliamentary proceedings, including the links between speeches, questions, answers, interjections, chamber context, and changing speaker careers over time.

This paper addresses that problem by introducing Hansard DB, a relational database of Australian federal parliamentary speech spanning Federation to the present. Hansard DB integrates speeches, questions, answers, and interjections with parliamentarian metadata from the Parliamentary Handbook in a single structured resource designed for flexible longitudinal and computational analysis. The database is built through a multi-stage parsing and validation pipeline intended not only to improve usability, but also to make visible where the underlying record is uncertain, inconsistent, or incomplete.

This paper makes three principal contributions to the study of Australian parliamentary discourse:

1. **Practical contribution**: We construct and openly release a comprehensive, cleaned, integrated, and queryable relational database of Australian Hansard proceedings from Federation (1901) to the present, including speeches, questions, answers, and interjections with full metadata from the Parliamentary Handbook.
2. **Technical contribution**: We document and release the full parsing pipeline, including automated detection and correction of date errors, speaker attribution issues, and service inconsistencies, providing a reproducible method for similar parliamentary data projects.
3. **Epistemic contribution**: We show that responsible large-scale analysis of parliamentary records requires confronting both technical errors

and systemic limitations. Understanding what Hansard records—and what it omits—is essential for valid interpretation.

The rest of this paper is structured as follows. The Background section provides necessary context on the Australian parliamentary system and explores how Hansard is constructed as a textual artefact. The Method section details the pipeline for transforming raw Hansard XML into a relational database, including the automated validation procedures and key design decisions. The Results section demonstrates Hansard DB's utility. Finally, the Discussion section explores the possible uses of this resource and the limitations of its interpretation.

Background

The Australian Parliamentary Context

Australia's Federal Parliament broadly mirrors other Westminster-style systems seen throughout various Commonwealth countries (Nethercote, 2016). It has two chambers: the House of Representatives and the Senate. House members represent constituencies, while senators represent the states (AEC, n.d.). The government is formed from the party or coalition with a majority in the House of Representatives (Ward, 2014). Unlike its British inspiration, the Australian Senate has strong powers to block legislation, leading some to describe the system as a 'Washminster' model: a blend of Westminster tradition and the powerful upper house characteristic of the United States Senate (Thompson, 2001). Australia's states and territories also have their own parliaments, which operate independently of the Federal Parliament within their respective jurisdictions.

The formal head of state is the monarch, represented in Australia by the Governor-General, who plays a largely ceremonial role in the legislative process (Parliament of Australia, n.d.-b). Day-to-day leadership rests with the Prime Minister, who heads a cabinet of ministers responsible for various governmental portfolios (Parliament of Australia, n.d.-c). The second-largest party or coalition in the House forms the opposition, led by the Leader of the Opposition and supported by shadow ministers in a shadow cabinet. The Opposition's role is to hold the government accountable, scrutinise its actions, and offer alternative policies to the public (Parliament of Australia, n.d.-c).

Procedural and Political Influences on Parliamentary Speech

Formal constraints on speech

Communication within Australia's parliamentary chambers is profoundly shaped by the procedural rules and customs that define the Westminster system. Parliamentary language is tightly regulated by standing orders, formal rules that dictate not only what may be said, but also how, by whom, and when (Parliamentary Education Office, 2026). Members must direct all comments through the Speaker, use respectful and non-inflammatory language, and avoid references to ongoing legal matters or personal attacks. These constraints enforce a formal and often indirect style of speech, where criticism must be carefully phrased and rhetoric is highly ritualised.

The daily routine of chamber business reinforces established norms. Proceedings follow a structured sequence: formal openings, consideration of government bills within strict procedural frameworks, and members' statements that offer brief opportunities to address issues beyond current bills (Ward, 2014). Questions without notice are a notable exception and often attract significant media and public attention (Parliament of Australia, n.d.-d). During Question Time, government backbenchers may pose friendly questions – known as Dorothy Dixers – while opposition members directly challenge ministers on virtually any aspect of executive activity, with the government expected to respond. Despite its ostensible purpose as a mechanism of government accountability, Question Time frequently serves to score political points rather than genuinely seek information or hold the executive to account (Hebden & Perche, 2023; Ilie, 2022).

Performance and party dynamics

Although parliamentary speech is structured around the principle of open debate, these exchanges are seldom spontaneous contests of policy or ideas (Bayley, 2004). Instead, debates in the Australian Parliament function as carefully choreographed performances, where rhetorical jabs and gestures are less about policy construction and more about signalling party strength, unity, and competence to the public. The image of adversarial contest is itself a political performance, designed to show that representatives are energetically advocating for their causes – even when the real policy bargaining happens off camera (Crewe, 2020; Denniss, 2025).

Political parties are central to this performance because they organise, coordinate, and control what unfolds in the chamber. Party strategy and policy formulation occur behind closed doors in party rooms – the "backstage" – while parliament serves as the "front stage" for a unified public presentation (Alasuutari, 2025; Karlsson et al., 2022). This backstage coordination is necessary because parties must present themselves as coherent collective actors if they are to compete effectively for office, maintain legislative unity, and govern successfully. It therefore constrains the policy choices and rhetoric available to individual MPs, who are expected to support the party's agreed position (Bowler et al., 1999). Party leaders maintain discipline by controlling members' career incentives and parliamentary opportunities – such as advancement, committee positions, access to party resources, and renomination – which reduces the scope for open dissent (Bowler et al., 1999; Proksch & Slapin, 2014).

Party management also orchestrates who speaks and what is spoken about at a broader level. Opportunities to speak are granted rather than intrinsic rights, and the business of the chamber is tabled and prioritised according to the government's agenda (Bäck et al., 2019; Ward, 2014). Within these longstanding procedural rules, speeches and exchanges become part of a broader political performance, serving not only parliamentary business but also the pursuit of political advantage and the management of public perception.

The recorded audience

Parliamentary speech is inherently mediated, addressed not only to parliamentarians participating in debate but also to broader publics watching from afar, especially the media through which it is reported (Davis, 2009a, 2009b; Vliegenthart et al., 2016). This mediation operates in two directions.

On one hand, parliamentarians perform both for those present in the chamber and for audiences beyond it: not only for the general public, but also to reassert leadership, signal to party members, frame key debates for their own side, and position themselves against opponents (Proksch & Slapin, 2014). Theatrical actions and visual stunts are often deliberately designed to resonate in the room while also lending themselves to media coverage and wider circulation (Edelman, 1988).

A particularly vivid illustration occurred in 2017, when then-Treasurer Scott Morrison brought a lump of coal into the House of Representatives as a

deliberate prop to make a political point about energy policy.

> Mr MORRISON (Cook—Treasurer) (14:28): This is coal. Do not be afraid. Do not be scared. It will not hurt you.
>
> The SPEAKER: The Treasurer knows the rule on props.[1]

This stunt, and others like them, are theatrical performances staged for audiences beyond the chamber itself – deliberate acts designed to be watched, recorded, and circulated to the broader public.

On the other hand, awareness of being recorded can shape parliamentary speech, encouraging strategic avoidance such as evading questions, qualifying answers, or withholding potentially damaging remarks (Ilie, 2022). Parliamentarians may also explicitly acknowledge this mediated context. For example, when MP Julian Hill told the House about his daughter's conception, he remarked:

> "I won't tell the story of how my daughter was conceived in the Hansard."[2]

Some MPs also push back against intensified recording and scrutiny, arguing that elected representatives need space to exercise judgement rather than simply perform for cameras (Chisholm, 2005; Smith, 2018). Their objection is not necessarily to publicity as such, but to forms of scrutiny that flatten parliamentary activity into easily consumable metrics or spectacles, without capturing the institutional constraints and contextual factors – such as party whipping, pairing arrangements, selective participation, illness, or leave – that shape parliamentary behaviour. In this view, increased visibility without increased public understanding can distort, rather than enhance, democratic accountability.

Hansard as an Edited Textual Artefact

Editorial decisions

While Hansard is often presented as a verbatim transcript, it is in fact a carefully edited textual artefact that both represents and reshapes parliamentary

[1] House of Representatives Hansard, 9 February 2017, p. 536
[2] House of Representatives Hansard, 23 March 2026, p. 161

speech (Edwards, 2016; Kotze et al., 2023; Mollin, 2007; Slembrouck, 1992). Reported in 2023 to comprise around 45 staff members (Coleman, 2023), Hansard's reporters and editors collectively determine what enters the official record. These staff first produce a "proof" transcript, published within hours of a sitting (Parliamentary Education Office, n.d.). A central part of the editing process is the removal of repetition and digression to produce a more coherent and readable text. Grammar and syntax are corrected, colloquialisms are smoothed out, and incomplete sentences may be clarified, all with the stated aim of capturing what the speaker meant to say rather than their exact words (Parliamentary Education Office, n.d.). In some instances, Hansard editors may remove entire passages or speeches from the official record if they are judged to be out of order, offensive, or in breach of parliamentary rules (Feldman, 2023; Hames et al., 2025). Members of Parliament may also request factual corrections to proofs, such as names or historical details, but may not alter the meaning or substance of proceedings. The final official Hansard then replaces the proof version, typically around two weeks later (Parliamentary Education Office, n.d.). The cumulative effect of these editorial decisions is to produce a text that is more formal, orderly, and controlled than the debate as originally spoken (Kotze et al., 2023).

Interjections from other members—such as jeers, cheers, or laughter—are also selectively recorded, often only when they are judged to constitute material interruptions rather than background noise, as determined by Hansard editorial policies (Edwards, 2016). This selectivity is compounded by the absence of any systematic attendance data, since Hansard provides no record of how many members were actually present in the chamber during a given debate. While formal quorum requirements must be met (40 members in the House of Representatives and 32 in the Senate) for a chamber to sit officially (Parliament of Australia, n.d.-e), the record does not distinguish between a minister addressing a full chamber and one speaking to only a handful of parliamentarians.

At the same time, this editorial apparatus is not static, and staff turnover, evolving guidelines, and individual variations in judgement mean that consistency across the corpus cannot be assumed. Judgements about what counts as a "material" interjection, whether a remark is sufficiently "offensive" to omit, or how best to render an incomplete sentence are inherently subjective and may vary from one parliament to the next, from one session to the next, or even from one day to the next. One reporter may regard a particular interjection as worth recording, while another, on a different day or in a different debate, may let it pass without comment. These difficulties

are compounded by the apparent absence of publicly available editorial policies, which makes it difficult to track how such judgements have been made, standardised, or revised over time (Edwards, 2016). This temporal variability is an inevitable feature of any human-mediated transcript produced over more than a century of continuous operation and introduces a further layer of editorial filtering beyond the formal procedural rules themselves.

What can never be captured

Beyond editorial choices, there are fundamental limitations inherent to the transformation from speech to text that no amount of editing can overcome. Elements such as tone, gesture, and emphasis are inevitably lost. The absence of vocal inflection, facial expression, and physical movements alters how statements are interpreted, often flattening emotion or ambiguities that would be apparent in person or in broadcast video (Bucholtz, 2000).

To give one example: when Tim Wilson, member for Goldstein, sung a verse of Billy Joel's "We Didn't Start the Fire" adapted to mock the Treasurer, Hansard recorded the lyrics but not the melody, producing a textual artefact that bears no trace of the performance's tone:

> You know, there's a Billy Joel song that sounds kind of relevant: 'The Treasurer did start the inflation fire. The inflation is burning while the Treasurer is squirming. The Treasurer did start inflation fire. Yes, he poured debt petrol on it, and he cashed organised crime to fuel it.'[3]:

Such examples illustrate the fundamental limitations of treating parliamentary transcripts as transparent records of communication – missing not just context, but entire modalities of expression.

The use of props and visual stunts, which frequently attract media attention (Mclellan, 2019), is rarely documented in official records, further limiting the full context of the communication. A particularly vivid illustration occurred in 2017, when Senator Pauline Hanson entered the Senate wearing a burqa (an Islamic face veil), a deliberately provocative act that generated significant media coverage. Yet the Hansard record captures only the procedural aftermath:

[3]House of Representatives Hansard, 4 March 2026, p. 45

> *Senator Hanson having entered the chamber–*
>
> Senator DUNIAM: What on earth?
>
> *Honourable senators interjecting–*
>
> The PRESIDENT: Senators, order! Senators, I've been advised by the clerk via the attendant that the identity of Senator Hanson was established before she entered the chamber. I'm just going to reflect on the mode of dress that Senator Hanson is using. We'll continue with question time.[4]

The Hansard transcript records the chaos but cannot convey the visual spectacle that dominated news coverage – the burqa itself, the reactions of other senators, or the theatrical dimension of the moment.

As a result, Hansard offers only a partial reflection of parliamentary communication: it provides transparency and accountability to the public, but omits much of the immediacy, spontaneity, and theatricality that characterise live debate.

Hansard is therefore not a transparent or complete record of parliamentary communication. In transforming speech into text, it necessarily strips away tone, gesture, visual performance, and other expressive elements that shape how speech is understood in the chamber and beyond. Yet as the continuous official record of what is formally said, it remains indispensable for analysing longer-term patterns in parliamentary discourse, including framing, agenda-setting, and the quality of debate. The value of a properly structured Hansard database lies not in overcoming the inherent limits of transcription, but in enabling this kind of careful longitudinal analysis at a scale that existing fragmented resources do not allow.

Method

The transformation of raw parliamentary transcripts into a structured, queryable relational database proceeds through a multi-stage pipeline. We begin by using custom scrapers, which automatically retrieve files from online sources, to collect raw XML transcripts from the official Parliament Hansard archive and from a pre-existing historical collection, which are then stored in the database. In parallel, we retrieve parliamentarian metadata, including

[4] Senate Hansard, 17 August 2017, p. 5980

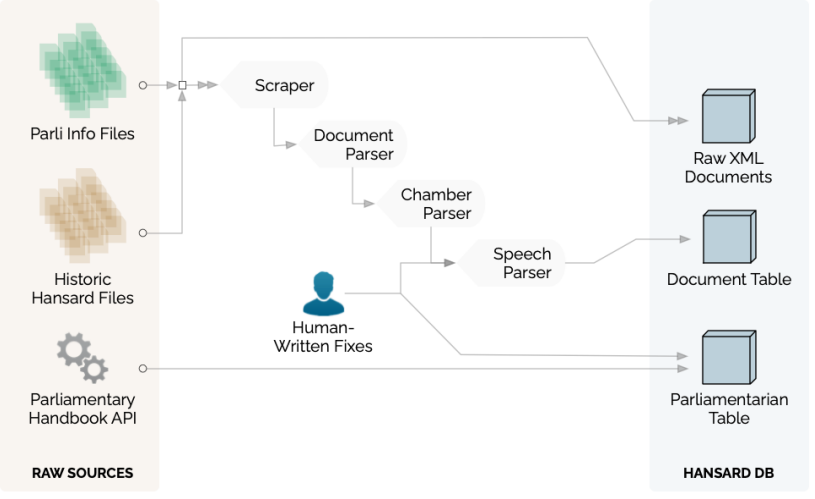

Figure 1: Overview of the multi-stage processing pipeline used to transform raw Hansard XML and Parliamentary Handbook metadata into a structured relational database. Arrows indicate data flow from acquisition through document-, chamber-, and speech-level parsing to final database storage, with automated validation and human-written fixes applied at the database stage.

names, party affiliations, electorates, and service histories, from the Parliamentary Handbook API, through which these records are made available in machine-readable form.

These two data sources form the inputs to the processing workflow. The raw XML documents are then processed in three stages by parsers, which extract and reorganise information from structured files. First, document-level parsing identifies high-level information such as the date, parliament number, session, and chamber, and separates each document into major structural blocks such as the main chamber or the Federation Chamber. Second, chamber-level parsing divides these blocks into individual contributions such as speeches, questions, and answers. Third, speech-level parsing converts the XML-formatted contributions into clean, structured text objects ready for database storage, while extracting the substantive content, speaker attribution, and associated metadata for each contribution.

Once these records have been stored in the database, two further automated processes are applied. First, internal joining logic links each speech to the relevant parliamentarian record by matching speaker identifiers to member records. Second, fixes logic applies manually specified corrections to resolve

known problems in either the joining stage or the parser output, including date errors, alternative parliamentarian identifiers, and inconsistencies in service periode errors, reconciling alternative parliamentarian IDs, or fixing service period discrepancies.

Data sourcing

The foundation of this project is the comprehensive acquisition of Australian parliamentary transcripts in machine-readable format, drawn from two principal sources: Tim Sherratt's hansard-xml repository [5] and documents scraped directly from the Parliament of Australia Hansard page [6] . Both the official parliamentary archive and the hansard-xml repository are available under Creative Commons licences [7].

Hansard-xml encompasses proceedings from 1901 to 2005, offering scraped XML documents alongside detailed documentation of missing sessions and record anomalies. Although hansard-xml represents a vital historical dataset, it is important to acknowledge that the official Hansard archive may be retroactively updated to incorporate corrections or additional material (Sherratt, 2024). Consequently, hansard-xml may occasionally lag behind the most current official version if amendments occur subsequent to the repository's latest update. Nonetheless, it was utilised due to the considerable challenges associated with alternative sources.

For debates from 2000 to the present, this project uses a scraper targeting the Parliament's Hansard ParlInfo interface. OpenAustralia was reviewed but not used as a primary source as it stores proofs, rather than the finalised documents.

By combining the extensive historical coverage of hansard-xml with up-to-date and authoritative extractions from the official parliamentary archive, this project assembles a unified Hansard corpus – from federation to the present day.

[5] https://github.com/wragge/hansard-xml
[6] https://www.aph.gov.au/Parliamentary_Business/Hansard
[7] https://www.aph.gov.au/Help/Disclaimer_Privacy_Copyright#c

Parsing

Document Level

The Hansard XML documents generally follow a standard structure, with a single <hansard> root containing a <session.header> and multiple section transcripts, as described in Figure 2.

```
<hansard ... >
    <session.header>
    </session.header>
    <section1.xscript>
    </section1.xscript>
    ...
</hansard>
```

Figure 2: Canonical root structure of a Hansard XML document. The <hansard> element contains a single <session.header> with sitting-day metadata and one or more <section.xscript> elements, each capturing a distinct set of proceedings (e.g. main chamber, Federation Chamber, or answers to questions on notice).

Each section captures a distinct set of proceedings – such as the main chamber, Federation Chamber, or answers to questions on notice – and was parsed separately. The <session.header> element contains metadata about each document, extracted from the following child tags: <date> (sitting date), <parliament.no> (parliament number), <session.no> (session number), <period.no> (period number), and <chamber> (House of Representatives or Senate).

Hanging debates – debate elements situated directly under the <hansard> root rather than within a designated section – required specialised handling. These elements were reclassified based on the occurrence of the term "notice" in their metadata: if "notice" and "answer" appeared in the title (excluding the plural form "notices"), the element was categorised as an answer to a question on notice; otherwise, it was assigned to the main chamber. Manual review was conducted to verify that this keyword-based approach produced valid results.

Chamber Level

Each <section.xscript> element was parsed independently, as these sections do not follow a strict format. Instead, they use a flexible hierarchy built

from <debate> elements, typically featuring an embedded <debateinfo>. Debates may also contain nested <subdebate.1> elements, which in turn can contain further <subdebate.2> elements, with each subdebate level including its own metadata (<subdebateinfo>). At every hierarchical level – debate or subdebate – the proceedings are ultimately expressed through <speech>, <question>, or <answer> elements, which are the primary elements of interest for this dataset. This structure is demonstrated in Figure 3.

```
<section.xscript>
    <debate>
        <debateinfo>
            <title>BILLS</title>
        </debateinfo>
        <speech>...</speech>
        <question>...</question>
        <subdebate.1>
            <subdebateinfo>
                <title>Specific Bill</title>
            </subdebateinfo>
            <speech>...</speech>
            <subdebate.2>
                <subdebateinfo>
                    <title>Amendment Discussion</title>
                </subdebateinfo>
                <speech>...</speech>
            </subdebate.2>
        </subdebate.1>
    </debate>
</section.xscript>
```

Figure 3: Representative <section.xscript> element illustrating the flexible, nested hierarchy used to organise chamber proceedings. <debate> elements may contain <subdebate.1> and <subdebate.2> children, each with its own <debateinfo> or <subdebateinfo> metadata block. Speeches, questions, and answers appear at any hierarchical level and are extracted in document order to preserve the flow of debate.

To extract these core elements, each <section.xscript> was systematically searched for any <speech>, <question>, or <answer> tags, preserving their sequential order as it appears in the document. This ensures that the parsed records faithfully represent the original chamber proceedings, capturing both the structure and flow of debate.

Question and Answer Pairings

One of the unique features of the Australian parliamentary process is the formal questioning of government, which is recorded in Hansard using

<question> and <answer> elements. However, the XML structure does not explicitly link individual questions to their corresponding answers, presenting a challenge for data extraction.

To address this, a pairing strategy was implemented: for each <question> element, the parser searches forward through its siblings within the same parent element to find the first <answer> element. If such an answer exists, it is paired with the question. This approach leverages the typical ordering and nesting of questions and their responses, allowing the dataset to reconstruct the logical connections between questions and answers, even in the absence of explicit identifiers. Questions without answers and answers without questions are both handled gracefully and stored as unpaired elements, ensuring no data is lost during processing.

Debate Information Extraction

To provide richer context for each speech, question, or answer, it is important to capture the titles and topics of the debates and subdebates in which they occur. Hansard embeds this contextual data within the elements: <debateinfo> and <subdebateinfo>. However, these informational elements may appear at different hierarchical levels within the chamber transcript.

To systematically extract hierarchical debate context, a recursive method was used. For each target element (such as a speech), the parser traverses up through the XML tree, moving from the current element to its parent. At each level, it searches for a directly nested metadata element – either <debateinfo>, <subdebateinfo>, or <title>. If a <title> tag is found, its text is cleaned to remove excessive whitespace and appended to a list of titles. If a <debateinfo> or <subdebateinfo> is found, the parser extracts the embedded <title> from within. The search proceeds upward through each ancestor until the root is reached or no further parents are found.

Once all titles from the current element up through its ancestral debates and subdebates have been collected, their order is reversed. This reversal ensures the final title string reflects the natural top-down progression (from broader debate context to narrower subdebate). The resulting string, with each level separated by commas, is then attached to each speech, question, or answer as a descriptive context label.

Speech Level

At the speech level, Hansard records are structured around `<speech>` elements (which also serve as the base for `<question>` and `<answer>` elements). Each speech element contains a speaker identifier (the talker), substantive content (text), and may contain any number of interjections (interruptions from other members).

The talker identifies who is speaking and is typically provided through a `<talker>` element containing a `<name.id>` unique identifier. The speech content is distributed across one or more `<para>` or `<p>` elements, which can appear as direct children of the speech element, within a `<talk.start>` block, or interspersed with interjections.

Interjections represent interruptions to a speech by another member. They may appear in different forms depending on the era: attributed interjections where the speaker is explicitly identified through nested `<talker>` elements within an `<interjection>` block, or inline interjections where the interruption is marked within the flow of the main speech text.

The goal of the speech level parsers is to extract from each XML element: the talker (speaker identifier), the substantive text of the speech, and any interjections along with their attributed speakers where applicable. Interjections were sorted into three types: - Office: Interjections made by an officer of the house, most often procedural interjections - General: Interjections where either the speaker or the words spoken were not recorded. - Speaker: Interjections where a specific part is attributed to a member or senator who is not an officer of the house. To deal with the varied format of the XML structure across the history of the digitised Hansard, nine distinct parser classes were created. These can be grouped into three major eras that represent major shifts in the XML structure: Early Digital (1981-1997), Mass Digitisation (1901-1980 + 1998-2011), and Modern (May 2011-Present).

Early Digital (1981-1997)

The Early Digital era contains three sub-parsers (hansard1981, hansard1992, hansard1997) which handle the transition from paper-based records to early digital formats.

Speaker identification differs across the sub-parsers. The 1981 parser extracts the speaker identifier from `<talker>/<name.id>` elements within

```
<SPEECH ... NAMEID="XXX" ... >
  <TALK.START>
    <TALKER>
      <NAME.ID>XXX</NAME.ID>
    </TALKER>
    <PARA IN-LINE="1">—Opening text...</PARA>
  </TALK.START>
  <PARA>Main speech content continues here...</PARA>

  <INTERJECT CHAIR="0" ... NAMEID="YYY">
    <TALK.START>
      <TALKER>...</TALKER>
      <PARA IN-LINE="1">—Interjection text...</PARA>
    </TALK.START>
  </INTERJECT>

  <PARA>—Members interjecting </PARA>

  <PARA>
    <EMPHASIS FONT-WEIGHT="BOLD">The SPEAKER<\EMPHASIS>Order!
  <PARA>


  <PARA>Speech continues after interjection...</PARA>
</SPEECH>
```

Figure 4: Representative <SPEECH> element from the Early Digital era (1981–1997). Speaker identification relies on <NAME.ID> within a <TALK.START> block (1981 sub-parser) or on a NAMEID attribute on the speech element itself (1992 and 1997 sub-parsers). Block-level interjections appear in dedicated <INTERJECT> elements with nested <TALKER> attribution; inline interjections are identified heuristically from short paragraphs containing "interjecting" or from bolded name spans at the start of a <PARA>.

a <talk.start> block (See Figure 4, line 4). The 1992 and 1997 parsers instead use a nameid attribute directly on the speech element itself, which considerably simplified the parsing (See Figure 4, line 1). Inline interjections had no usable speaker information.

Interjections in this era appear in two distinct forms. Block-level interjections are contained within dedicated <interject> or <interjection> elements, which include their own <talker> sub-elements for speaker attribution (See Figure 4, line 10-15). Inline interjections in this format are poorly identified, hence a set of heuristics are required to parse them. Short paragraph elements with the word "interjecting" are used to identify general interjections (See Figure 4, line 17). Similarly, short, fully italicised elements are also used to signify a general interjection. For speaker and office interjections, it was found that a bolded name at the start of the para element

indicated an interjection (See Figure 4 line 19-21).

For Block interjections, they were found to always be speaker or office interjections. To differentiate, the presence of `chair="1"` or `nameid=10000` (the placeholder for office speakers) was used to identify office interjections, with all the rest assumed to be speaker. For inline cases, the parser identifies `<emphasis>` elements with `font-weight="BOLD"` for office roles or `font-slant="ITAL"` for general interjections. There were no found examples where an inline interjection was a speaker interjection.

Mass Digitisation (1901-2011)

The Mass Digitisation era spans contains three sub-parsers (hansard1901, hansard1998, hansard2000). This era introduces a more structured approach to XML formatting, with detailed `<talker>` metadata, but lacks the `<talk.text>` container found in the Modern era (See Figure 5, line 23).

Speaker identification in the parser first checks `<talk.start>/<talker>/<name.id>` (See Figure 5, line 3), falling back to `<talker>/<name.id>` if not found. It also examines `<continue/talk.start/talker/name.id>` for alternate speakers for cases where the speech is started by an office holder, but then continued by the speaker (See Figure 5, line 32). A special case handles scenarios where an office holder (`nameid="10000"`) starts the speech with an interjection, which is often not recorded within an interjection element.

Interjection identification in this era employs multiple strategies: explicit `<interject>` or `<interjection>` elements (See Figure 5, line 14); `<talk.start>` or `<continue>` elements where the speaker has the id `10000` (office holder); and inline interjections identified through bold or italic `<inline>` elements within `<para>` elements (See Figure 5, lines 10-19). Various heuristics were designs to `<inline>` elements that signify interjections, rather than names of places, bills or other descriptive commentary.

Block interjects were always identified as either office or speaker interjections, relying on the `name.id` to determine the type. Starting from 2000, it was possible for a block interjection to not include any text, indicating a general interjection made by the identified speaker. Inline interjections could be any type of interjection. Bolded inline elements at the start of a para indicated a change of speaker, and were then checked for the presence of office keywords (SPEAKER, CLERK, PRESIDENT, CHAIR, DEPUTY) to see if they were office interjections, if they were not, they were seen as speaker

```xml
<speech>
  <talk.start>
    <talker>
      <name.id>XX4</name.id>
    </talker>
    <para>—Opening speech text...</para>
  </talk.start>
  <para>Main speech content continues...</para>

  <para>
    <inline font-weight="bold">OTHER MP</inline>
    interjects from across the chamber...
  </para>

  <para>
    <inline font-style="italic">
      interjection description
    </inline>
  </para>

  <interjection>
    <talk.start>
      <talker>
        <name.id>10000</name.id>
      </talker>
      <para>—Office interjection text...</para>
    </talk.start>
  </interjection>

  <continue>
    <talk.start>
      <talker><name.id>XX4</name.id></talker>
      <para>—Speech continues...</para>
    </talk.start>
  </continue>
</speech>
```

Figure 5: Representative <speech> element from the Mass Digitisation era (1901–1980 and 1998–2011). Speaker identification uses <talk.start>/<talker>/<name.id>, with fallback to <continue> elements for speeches begun by an office holder. Block interjections appear in <interjection> elements; inline interjections are identified from bold or italic <inline> elements within <para> nodes. Unlike the Modern era, this format lacks a <talk.text> container to bound speech content.

interjections (See Figure 5, lines 10-13). Italicise inline elements were found to be general interjections (See Figure 5, lines 15-19).

Block interjections consistently had successful talker attribution through the `talker.name.id` elements. For inline interjections, talker attribution was not possible.

Modern (May 2011-Present)

The Modern era uses a completely restructured XML format resembling HTML, with three sub-parsers (hansard2011, hansard2012, hansard2021) that share the same base class with limited differences between them. This era represents the most structured approach to parliamentary record-keeping, introducing the `<talk.text>` container to bound speech content (See Figure 6, line 23).

Speaker identification follows a consistent pattern across all sub-parsers: `<talk.start>/<talker>/<name.id>` (See Figure 6, line 3). The structure is uniform and well-defined.

All interjections in this era are inline, identified through `<span>` elements with specific class attributes (See Figure 6, line 14-20):

- `HPS-OfficeInterjecting` (Office Interjection)
- `HPS-OfficeContinuation` (Office Interjection)
- `HPS-OfficeSpeech` (Office Interjection)
- `HPS-MemberIInterjecting` (General, known speaker, unknown words)
- `HPS-GeneralIInterjecting` (General)
- `HPS-MemberInterjecting` (Speaker)
- `HPS-GeneralInterjecting`. (General, known words, unknown speaker)

The interjection type is determined directly by the span class – office, member, or general – with additional logic checking for office role keywords (SPEAKER, DEPUTY, CLERK, PRESIDENT, CHAIR) within `HPS-MemberSpeech` spans.

Text extraction for normal speeches only pulls content from `<span>` elements with class `HPS-Normal` or `HPS-Small` (See Figure 6, lines 9-11). For inline interjections, the extraction logic varies by class: `HPS-MemberIInterjecting` and `HPS-GeneralIInterjecting` extract only the description text, while `HPS-GeneralInterjecting` extracts the text plus any tail content.

```
<speech>
  <talk.start>
    <talker>
      <name.id>XXX</name.id>
    </talker>
  </talk.start>
  <talk.text>
    <body>
      <p class="HPS-Normal">
        <span class="HPS-Normal">
          <a href="XXX" type="MemberSpeech">
            <span class="HPS-MemberSpeech">Mrs Jones</span>
          </a> (Grayndler):  Opening speech...
        </span>
      </p>
      <a href="YYY" type="MemberSpeech">
          <p class="HPS-Normal">
            <span class="HPS-Normal">
              <span class="HPS-MemberIInterjecting">
                 Ms Smith interjecting
              </span>—
            </span>
          </p>
      </a>
      <p class="HPS-Normal">
        <span class="HPS-Normal">
          <span class="HPS-GeneralIInterjecting">Members interjecting</span>
        </span>
      </p>
      <p class="HPS-Normal">
        <span class="HPS-Normal">
          <span class="HPS-OfficeInterjecting">The SPEAKER:</span>  Order!
        </span>
      </p>
    </body>
  </talk.text>
</speech>
```

Figure 6: Representative <speech> element from the Modern era (May 2011–present). Speech content is bounded within a <talk.text>/<body> container. All interjections are inline, distinguished by class attributes: HPS-MemberIInterjecting (attributed member, words unrecorded), HPS-GeneralIInterjecting (unattributed), HPS-GeneralInterjecting (words recorded, speaker unattributed), and HPS-OfficeInterjecting/HPS-OfficeContinuation/HPS-OfficeSpeech (presiding officers). Speaker identifiers are extracted from href attributes on <a> elements, with a lookup table used when the link is absent after the first occurrence.

A key distinction from earlier eras is that inline authors are attributed. The parser extracts speaker identifiers from `<a>` elements with `href` attributes linking to the parliamentary database (See Figure 6, line 16). For cases where no `href` is present, a lookup table maps name text to `href` IDs, maintaining attribution even when the link is not directly embedded. This is required as the `href` will only appear for the first instance of a speaker in a speech.

Parliamentary handbook integration

While Hansard transcripts include basic speaker metadata, the way this information is recorded is often inconsistent. Across decades of parliamentary debate, members may appear under abbreviated names, formal or honorific titles, or variant spellings, and seat information is subject to change or ambiguity. These inconsistencies pose significant challenges for reliable author identification, especially in large-scale or longitudinal research.

The value of the Parliamentary Handbook API for overcoming these obstacles was demonstrated by Leslie (2024), who showed how the Handbook can be programmatically scraped to obtain comprehensive, structured metadata about all members of parliament. Following this approach, we ensure accuracy and consistency in our dataset by anchoring each Hansard record to the unique identifier encoded in the `<name.id>` field of the XML transcript. This persistent identifier serves as a stable key for joining Hansard entries directly to authoritative records from the Parliamentary Handbook API [8].

For each parliamentarian, we extract and store rich biographical and service information, including:

- A unique member ID
- Preferred, family, and middle names, as well as any known display variations
- Gender, First Nations status, date of birth, and image links
- Complete histories of party affiliation and representation by seat or state, including service periods in the House or Senate
- Chronological records of ministerial and shadow ministerial appointments

[8] https://handbookapi.aph.gov.au/api

With this process, almost every speech, question, answer, or interjection in the Hansard database is not only attributed to its textual author, but can be unambiguously linked to a detailed, structured record of the speaker's career. This integration ensures that, even across periods of name changes, shifting party allegiances, or repeated names, each parliamentary contribution is reliably connected to its correct author – enabling more precise, meaningful, and comprehensive research.

Database Format

While calls for a public-facing Hansard API (Angus, 2026) emphasise the value of flexible, large-scale querying, we argue that a local relational database offers many of the same advantages. Modern relational databases provide complex querying capabilities, rapid access, and seamless integration with analytical tools – paralleling much of the functionality of dedicated APIs, but without the need for centralised hosting or reliance on external infrastructure. By maintaining the parsed Hansard data in a standardised relational database, users can efficiently search, filter, and retrieve records, unconstrained by web-based interfaces.

Below, and illustrated in Figure 7, we outline the structure of the Hansard database, which is designed to maximise accessibility and analytical flexibility.

Parsed Documents

The scraping and parsing process is controlled by rows within the *Source* table, which links to multiple *RawDocument* records – each storing the original XML (or text), metadata about the extraction process, and a flag indicating whether the document is a proof. After parsing, each *RawDocument* is divided into multiple *SittingDay* entries, each representing a unique combination of house, date, and section (with section referring to Chamber, Federation Chamber, or Answers to Questions).

The core parliamentary contributions – speeches, questions, and answers – are structured as *Document* records. Each *Document* stores the text, title, type, and a reference to the author. Documents are self-referential: relationships such as *citedBy* and *references* enable questions to be linked to their answers, and vice versa. *Interjection* records are directly associated with

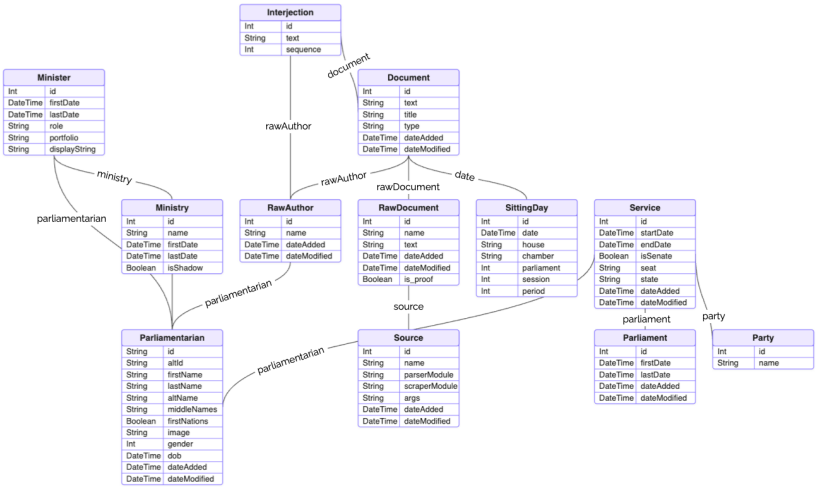

Figure 7: Entity–relationship diagram of the Hansard relational database. Core parliamentary contributions (*Document*, *Interjection*) are linked to sitting-day provenance (*SittingDay*, *Source*, *RawDocument*) and to speaker metadata (*Parliamentarian*, *Service*, *Party*, *Minister*, *Ministry*) via persistent Parliamentary Handbook identifiers. Self-referential *Document* relationships capture question–answer pairings.

their parent *Document*, storing their text, author, and sequence indicator for precise mapping to `[INTERJECTIONXX]` placeholders within the speech text.

Metadata

The core table of the metadata component is *Parliamentarian*, which contains a record for each unique member of parliament and is anchored by authoritative parliamentary handbook IDs. *Parliamentarian* entries include names, gender, date of birth, First Nations status, and alternative author IDs that occasionally appear in Hansard but are not canonical.

Parliamentarian records link to sets of *Service* records, which capture distinct periods of service as unique combinations of parliament, seat, and party affiliations. By merging service and party data from multiple official sources, this table simplifies downstream analysis.

In addition, *Parliamentarian* entries may be linked to *Minister* records, which detail any ministerial or shadow ministerial appointments. Each

Minister is associated with a *Ministry*, encompassing information about both ministries and shadow ministries, their leaders, and their timeframes.

Validation and corrections

Because Hansard XML varies substantially across time and contains known inconsistencies, we treated validation as a distinct stage of the pipeline. We aimed not simply to produce plausible parsed output, but to assess whether the parser preserved the structure and content of the source faithfully across the full temporal extent of the corpus. We combined parser-level checks on a targeted sample of files with post-ingestion checks across the full database. Where we identified clear and recoverable errors, we corrected them conservatively and documented them transparently.

Validation design

We conducted parser-level validation on 127 XML files selected across the full temporal range of the corpus to represent the major parser eras and structural formats in the Hansard archive. We parsed each file with the version appropriate to its date and converted the output to a normalised JSON representation for consistent inspection. Across this validation corpus, the parser extracted 8,250 parliamentary contributions and 12,332 interjections, which formed the basis for automated diagnostics and manual review.

After loading records into the relational database, we validated the full Hansard corpus at the post-ingestion stage. At this stage, we focused on corpus completeness, speaker attribution, question-answer reconstruction, and metadata consistency.

Pre-ingestion validation

We assessed parsed outputs with a suite of automated diagnostics designed to flag suspicious structures, including titles, timestamps, and malformed interjection markers appearing in parsed content. We then reviewed flagged cases manually against both the source XML and the rendered Hansard output. In addition, we read one debate from each validation file by hand to assess whether the parser preserved speaker sequence, interjection placement, and debate continuity.

We found no anomalies attributable to parser corruption. Instead, the irregularities we observed already existed in the source XML, including malformed markup, inconsistent attribution cues, unpaired question-answer structures, and procedural material embedded within speech elements.

Post-ingestion validation

Post-ingestion checks showed that corpus coverage is effectively complete, with more than 99.99% of sitting days for both chambers represented in the database (Figure Figure 8). The small number of missing days reflects rare gaps in the official archive, inaccessible or malformed records, or isolated historical anomalies.

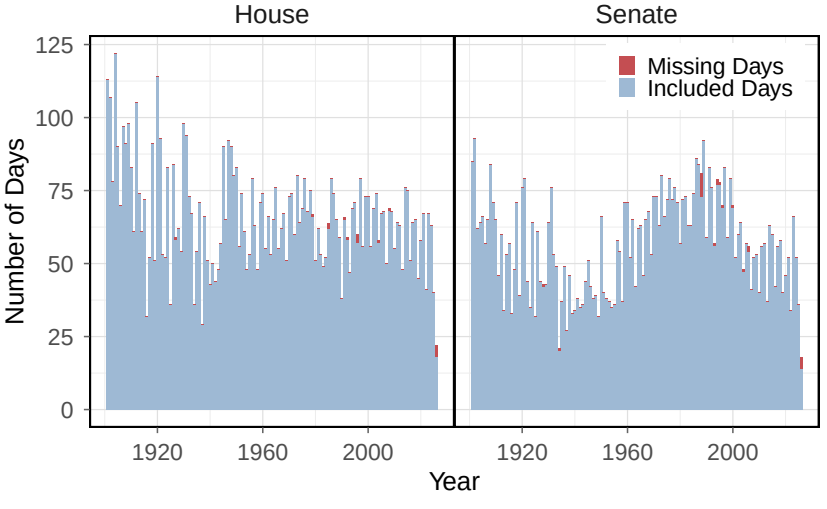

Figure 8: Coverage of sitting days in Hansard DB, 1901–present. Stacked bars show the number of sitting days included in the database (blue) and missing from the database (red) for each year, faceted by chamber. Missing days result from gaps in the official archive, inaccessible or malformed archival records, or historical anomalies. Overall coverage exceeds 99.99% for both chambers across the full temporal extent of the corpus.

We examined contributions without attributed speakers. These were concentrated in mass-digitised historical documents and were mainly short inline interjections. In many cases, the source text provides a name, but not in a form that can be reliably matched back to a unique speaker identifier. As Figure Figure 9 shows, these unresolved attributions are concentrated in

earlier material and reflect limitations in the source documents rather than systematic parsing errors.

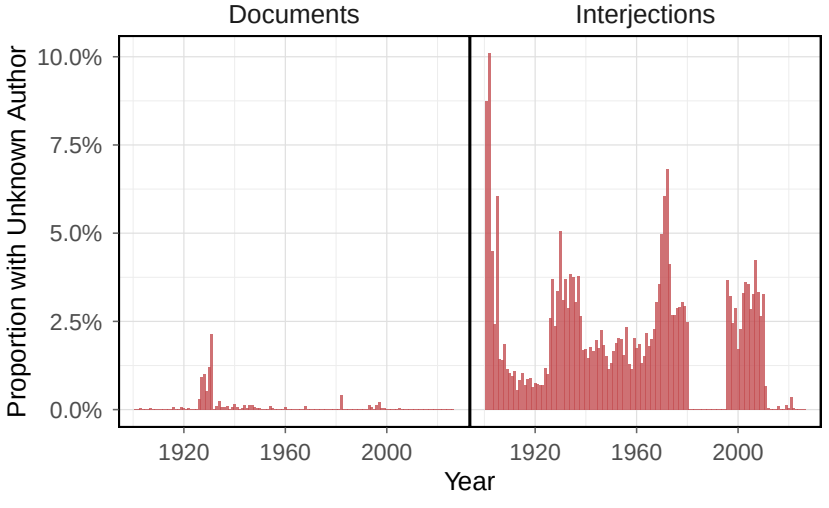

Figure 9: Proportion of parliamentary contributions with unresolved speaker attribution, by year. Left panel shows the annual proportion of speeches, questions, and answers for which the author field is blank or unknown; right panel shows the equivalent proportion for speaker-attributed interjections (type = 1). Elevated rates in earlier decades reflect sparse or inconsistent name identifiers in Mass Digitisation era XML documents (approximately 1901–2011), where inline interjections frequently lack usable talker metadata.

We evaluated question-answer reconstruction after ingestion. Table Table 1 and Figure Figure 10 show that unmatched questions and answers remain relatively uncommon and generally decline over time. Our manual inspection indicates that these mismatches usually reflect source-record irregularities, such as missing answers or misuse of question and answer blocks, rather than systematic parser failure.

Table 1: Summary of Question and Answer Pairs

Category	Count
Question-Answer Pairs	360440
Lone Questions	15082
Lone Answers	9366

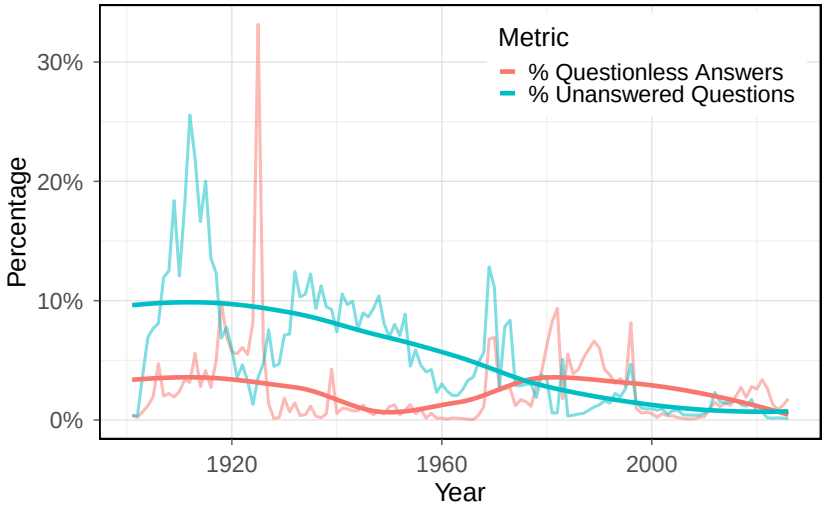

Figure 10: Annual rates of unmatched questions and answers, 1901–present. Lines show the proportion of questions with no matched answer (unanswered questions) and the proportion of answers with no matched preceding question (questionless answers) for each year; smoothed trends (LOESS) are overlaid. Both rates remain low and generally decline over time, indicating that unmatched pairs reflect source-record irregularities rather than systematic parser failure.

We then carried out a further corpus-wide attribution check against Parliamentary Handbook metadata. This check focused on cases where speeches could not be linked to an author, as well as cases where a speech was attributed to a person who, according to the handbook, was not serving in parliament on the relevant date. These checks helped identify unresolved attribution problems, distinguish genuine data issues from handbook inconsistencies, and isolate cases requiring manual verification.

Corrections

Corrections were applied only where the available evidence made the original record clearly wrong and the correct resolution could be established confidently. The aim of correction was therefore limited and conservative: not to reconstruct parliamentary proceedings beyond the evidence of the source, but to resolve egregious and verifiable errors while preserving fidelity to the published Hansard record.

One class of such corrections involved document dates. Each Hansard document should correspond to a unique sitting day and chamber, enforced

by a unique key on the `SittingDay` table. During validation, duplicate key conflicts revealed two cases where the date recorded in the XML was incorrect. In both instances, the correct date was confirmed by consulting the official PDF and then updated in the database.

A second class involved speaker attribution. Where malformed XML, typographical errors, or inconsistent identifiers prevented successful matching to a parliamentarian, the attribution was corrected only if the speaker could be established confidently from the official record. This process identified 132 alternative speaker identifiers used for 125 parliamentarians, all of which were reconciled to canonical handbook records. It also led to 31 speech reassignments after manual verification. In addition, we added 23 legitimate exceptions to an ignore list for identifiers that are not expected to link to a parliamentary member record, including the Speaker and foreign dignitaries, thereby preventing spurious warnings.

A third class involved metadata corrections arising from handbook integration. Because the Parliamentary Handbook serves as an external authority for service history, its inconsistencies can generate false attribution failures if accepted uncritically. These were corrected only where corroborating historical evidence such as parliamentary biographies was available, and all such changes were documented in the database comments for transparency and traceability. In total, this resulted in 6 seat corrections and 15 party corrections.

Overall, the remaining anomalies in the dataset should be understood primarily as source errors rather than parser errors. The parser was deliberately designed to fail conservatively: where the XML did not provide enough information to recover structure unambiguously, that ambiguity was preserved in the output rather than resolved through speculative inference. This design prioritises faithfulness to the published source over unwarranted reconstruction.

Examples of Hansard DB in use

The core product of this work is the reconstruction of Hansard into a database format – Hansard DB. In this section we demonstrate how the relational structure opens new possibilities for studying Australian parliamentary discourse – questions that were previously impossible to answer with longitudinal data at this scale. By integrating detailed biographical metadata – including gender, party affiliation, and parliamentary service periods – with

every parliamentary contribution, Hansard DB enables systematic analysis of how speaking patterns vary across the composition of parliament. Three example queries illustrate this capability: tracking how speech length has changed over time, measuring the concentration of speaking time across political parties, and investigating whether female parliamentarians receive more interjections than their male colleagues.

Are Speeches Getting Longer?

Speech length is more than a procedural detail: changes in its distribution can indicate shifting parliamentary norms, reforms to debate rules, or tighter regulation of speaking time (Bäck & Debus, 2016; Proksch & Slapin, 2014). Figure 11 calculates annual 50th, 75th, and 95th percentiles of speech length by chamber, allowing us to distinguish broad change from change concentrated among especially long speeches. Figure 12 visualises the results.

```sql
SELECT
    EXTRACT(YEAR FROM sd.date)::integer AS year,
    sd.house,
    PERCENTILE_CONT(0.50)
        WITHIN GROUP (ORDER BY LENGTH(d.text))::integer AS p50,
    PERCENTILE_CONT(0.75)
        WITHIN GROUP (ORDER BY LENGTH(d.text))::integer AS p75,
    PERCENTILE_CONT(0.95)
        WITHIN GROUP (ORDER BY LENGTH(d.text))::integer AS p95,
    COUNT(*)::integer AS n_speeches
FROM "Document" d
JOIN "SittingDay" sd ON sd.id = d."sittingDayId"
WHERE d.type = 'speech'
  AND d.text IS NOT NULL
  AND d.text != ''
GROUP BY year, sd.house
ORDER BY year, sd.house
```

Figure 11: SQL query computing the 50th, 75th, and 95th percentiles of speech length (in characters) for each year and chamber. Only non-empty speeches (`type = 'speech'`) are included. Results are used to produce Figure 12.

The results suggest that change is concentrated in the upper tail of the distribution. The 50th and 75th percentiles remain fairly stable over time, while the 95th percentile declines substantially. This shows how the database can be used to test claims that would be obscured by averages alone, and

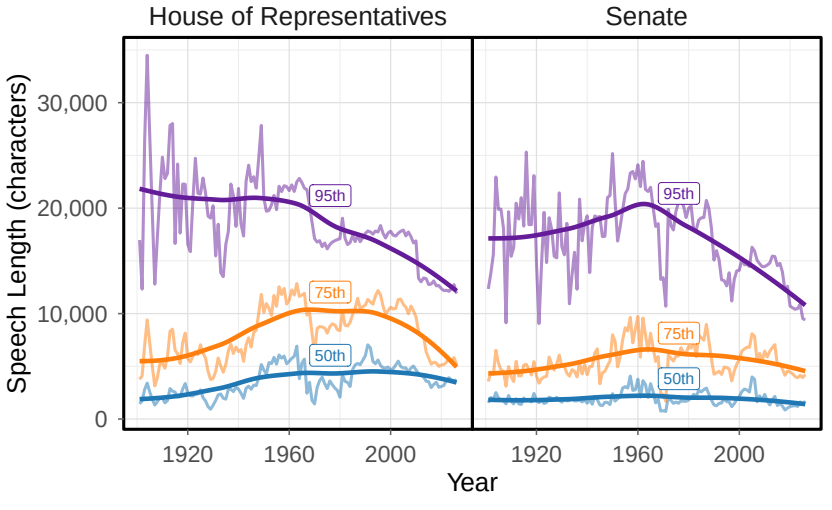

Figure 12: Distribution of speech length in the Australian Parliament, 1901–present. Lines show annual 50th, 75th, and 95th percentiles of speech length (characters) for the House of Representatives and Senate. Smoothed trends (LOESS) are overlaid on raw annual values. Panels are faceted by chamber.

to move from anecdotal impressions about "shorter speeches" to a more precise distributional account (Tucker et al., 2020).

This also opens up clear next steps. The same approach could be used comparatively across legislatures, or linked to questions about participation, agenda-setting, and deliberative quality. In that sense, speech-length distributions are not just descriptive outputs, but a way of evaluating broader theories about how parliamentary debate changes over time, including the effects of interventions and rule changes.

Is Parliamentary Speech Becoming More (or Less) Concentrated?

Who speaks is as important as how much is said. Democratic theory and empirical work on parliamentary representation both suggest that parliamentary voice should be broadly distributed rather than concentrated among a small number of parties (Bäck et al., 2019; Urbinati, 2006). Measuring concentration over time therefore provides a way to test such claims directly.

To do this, Figure 13 calculates the Herfindahl–Hirschman Index (HHI) of

party concentration in annual speech output for each chamber. As shown in Figure 14, higher values indicate that speaking time is dominated by fewer parties, while lower values indicate a broader distribution of parliamentary voice.

The results in Figure 14 show different trajectories across the two chambers. Senate HHI declines over time, indicating that speaking time has become less concentrated and more broadly shared across parties. House HHI remains comparatively stable, suggesting continued concentration among the major parties. In this way, the database makes it possible to evaluate claims about representation and party dominance longitudinally, rather than relying on anecdote or isolated cases.

These patterns also point to clear next steps. The same approach could be extended comparatively across legislatures, or linked to questions about legislative effectiveness, deliberative quality, and political trust.

Are interjection rates impacted by gender?

Whether women receive equal voice in deliberative institutions is a long-standing concern in democratic theory and gender-and-politics research. Role congruity theory suggests that women may face penalties for assertive participation, while deliberative scholarship emphasises that institutions shape whose contributions are treated as worth engaging (Eagly & Karau, 2002; Karpowitz & Mendelberg, 2014). Comparative research on parliamentary interruptions has produced mixed findings, but increasingly suggests that patterns of interjection can be used to test broader claims about gendered inclusion and exclusion in debate (Miller & Sutherland, 2023).

To examine this, Figure 15 calculates the mean number of interjections received per speech by year, chamber, and speaker gender. The results, shown in Figure 16, should be interpreted cautiously for women in the early decades of the series, as female representation before roughly 1943 is extremely sparse.

Across the full temporal extent of the corpus, female MPs consistently receive fewer interjections per speech than their male counterparts, aligning with the findings of Katz et al. (2026). This pattern is more plausibly read as a form of disengagement than as respectful restraint: male colleagues appear less likely to contest or engage with women's contributions on the floor,

```sql
WITH party_time AS (
    SELECT
        EXTRACT(YEAR FROM sd.date)::int AS year,
        sd.house,
        py.name AS party,
        SUM(LENGTH(d.text)) AS total_chars
    FROM "Document" d
    JOIN "SittingDay" sd ON sd.id = d."sittingDayId"
    JOIN "rawAuthor" ra ON ra.id = d."rawAuthorId"
    JOIN "Parliamentarian" p ON p.id = ra."parliamentarianId"
    JOIN "Service" sv
        ON sv."parliamentarianId" = p.id
       AND sd.date >= sv."startDate"
       AND (sd.date <= sv."endDate" OR sv."endDate" IS NULL)
    JOIN "Party" py ON py.id = sv."partyId"
    WHERE d.type = 'speech'
      AND d.text IS NOT NULL
      AND d.text <> ''
    GROUP BY 1, 2, 3
),
shares AS (
    SELECT
        year,
        house,
        total_chars::float
        / SUM(total_chars) OVER (PARTITION BY year, house) AS share
    FROM party_time
)

SELECT
    year,
    house,
    SUM(share^2) AS hhi
FROM shares
GROUP BY year, house
ORDER BY year, house
```

Figure 13: SQL query computing the Herfindahl–Hirschman Index (HHI) of speaking-time concentration by party for each year and chamber. Party shares are calculated as each party's total character count divided by the annual chamber total; HHI is the sum of squared shares. Only speeches with a matched service record are included. Results are used to produce Figure 14.

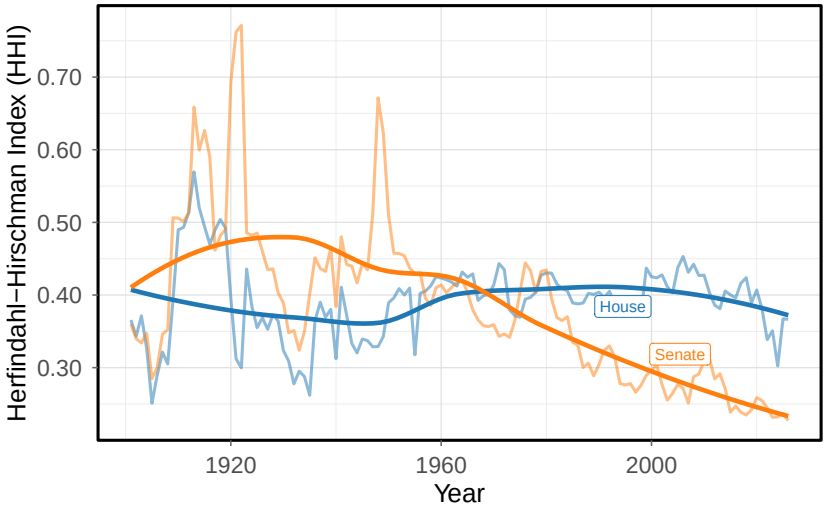

Figure 14: Party concentration of parliamentary speaking time, 1901–present. The Herfindahl–Hirschman Index (HHI) measures the degree to which speech is dominated by a small number of parties; higher values indicate greater concentration. Lines show annual HHI for the House of Representatives and Senate; smoothed trends (LOESS) are overlaid.

meaning that fewer interjections may signal a less contested – and therefore potentially less influential – parliamentary voice rather than a safer one.

Future questions are numerous. One would be to move beyond counts and analyse the content or tone of interjections; another would be to compare these dynamics across legislatures. More broadly, Hansard DB provides a way to connect parliamentary speech data to wider debates about gender, representation, and deliberative inclusion.

Discussion

The preceding results illustrate only a fraction of what Hansard DB makes possible. With over a century of parliamentary speech linked to biographical metadata, researchers can trace how speaking patterns, interjection rates, and the distribution of voice evolve across governments, parties, and historical periods. Recent advances in natural language processing – and in particular, large language models – expand these possibilities further: tasks like argument mining, stance detection, and rhetorical framing that were once intractable at scale are now increasingly feasible with this dataset (Angus, 2026).

```sql
SELECT
    EXTRACT(YEAR FROM sd.date)::int AS year,
    sd.house,
    p.gender AS gender,

    COUNT(*) AS n_speeches,
    COUNT(i.*) AS n_interjections,

    ROUND(
        (COUNT(i.*)::float
        / NULLIF(COUNT(*), 0)
        )::numeric,
        3
    ) AS interjections_per_speech

FROM "SittingDay" sd
JOIN "Document" d ON d."sittingDayId" = sd.id
LEFT JOIN "Interjection" i ON i."documentId" = d.id
JOIN "rawAuthor" ra on ra.id = d."rawAuthorId"
JOIN "Parliamentarian" p on p.id = ra."parliamentarianId"
WHERE d."type" = 'speech'
    AND sd."chamber" = 'Primary Chamber'


GROUP BY 1, 2, 3
ORDER BY 1, 2, 3
```

Figure 15: SQL query computing the mean number of interjections received per speech for each year, chamber, and speaker gender. Only primary-chamber speeches (`chamber = 'Primary Chamber'`) are included; interjections are joined via a left join to preserve speeches with zero interjections. Results are used to produce Figure 16.

Yet the capacity to ask new questions does not guarantee the capacity to answer them well. This section examines the epistemic boundaries of the Hansard DB: what it can reliably tell us, what remains uncertain, and where automated analysis risks producing confident but misleading conclusions.

Technical Limitations

Several epistemic limitations stem from the technical choices made in constructing Hansard DB, as well as exiting issues in the transcribed record.

First, a small number of sitting days are missing from the corpus. These gaps result from rare archival failures, inaccessible records, or historical anomalies. While the number is small – less than 0.01% of sitting days – it introduces uncertainty into any longitudinal pattern, particularly for periods

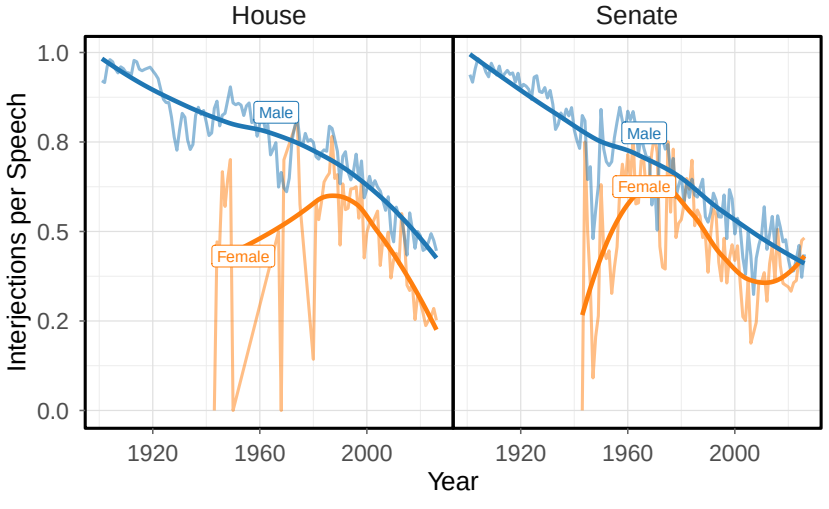

Figure 16: Interjections received per speech by gender, 1901–present. Lines show annual mean interjections per speech for male and female parliamentarians in the House of Representatives and Senate; smoothed trends (LOESS) are overlaid. Female representation in the federal Parliament began in 1943 and early conunts are based on very small speaker counts and exhibit high year-to-year variability. Panels are faceted by chamber.

where gaps are concentrated (see Figure 8).

Second, speaker attribution is not always complete. Some speeches lack sufficient information to unambiguously identify the author, and remain unmatched to a parliamentarian. These unmatched speeches are concentrated in the early (pre-1980) digitised records, where name formats were inconsistent and metadata was sparse. The effect is epistemic: analyses of individual speaking patterns, party distribution, or gender differences are based on a systematically incomplete subset for earlier decades.

Third, interjection parsing is imperfect. Heuristics were required to identify inline interjections placed within paragraph elements rather than in their own structural nodes. Without understanding the semantic content of the text, it is impossible to reliably distinguish these from other speaker references. The result is epistemic: interjection counts for earlier decades may systematically undercount certain types of interruptions, affecting any analysis of interactivity over time.

A deliberate design decision compounds these issues: Hansard DB was constructed to reflect Hansard as published, not to reconstruct what actu-

ally occurred in Parliament. Errors in the original record – misattributed speeches, incorrect dates, or garbled transcriptions – are preserved unless they were obvious enough to detect through rule-based validation. For speaker attribution specifically, while clearly erroneous matches were corrected (such as speeches attributed to members not yet elected), there will almost certainly be incorrect attributions that remain undetected because they do not violate any validation rules. Hansard DB inherits Hansard's errors alongside its content.

An important qualification, however, is that many of these limitations need not be fatal for downstream quantitative analysis if they are broadly systematic. Stable under-attribution, consistent editorial selection, or persistently imperfect interjection parsing may still permit valid comparative inference, particularly where the goal is to track relative rather than absolute differences. The greater technical risk arises when such imperfections vary over time in ways that are not observed in the data. Changes in transcription conventions, editorial workflows, annotation rules, or staffing can introduce structural breaks that appear analytically as shifts in parliamentary behaviour when they instead reflect changes in corpus production. This points to an important avenue for future research: treating Hansard not only as a record of Parliament, but also as the output of an evolving technical pipeline. Archival study of transcription procedures, editorial guidelines, and staffing transitions could help identify when discontinuities in the corpus reflect changes in speech and when they reflect changes in recording and annotation practice.

Interpretive Risks

Beyond technical limitations, Hansard DB inherits constraints from the source material that cannot be resolved through data cleaning or parsing improvements.

Most fundamentally, Hansard is not a verbatim record of chamber proceedings but an official textual rendering of them. It should therefore not be treated as identical to what happened in Parliament. As discussed in the Background, the transcript is shaped by editorial judgement and by the inherent limits of transcription itself. Features that may shape both meaning and political effect – including tone, pace, timing, hesitation, overlap, gesture, and audience response – are largely absent from the record. For analysis, this means Hansard DB is well suited to studying patterns in recorded

language, but less able to capture the full interactional and performative character of parliamentary speech.

This also limits what can be inferred about political significance. In the textual record, consequential interventions may appear little different from routine contributions. Hansard DB can show what was said, but not reliably which remarks landed, which fell flat, or which altered the course of debate. Analyses concerned with decisive moments, political impact, or the atmosphere of debate therefore require corroboration from sources beyond Hansard itself, such as broadcast footage, news reporting, memoir, or archival material.

Risks of Automated Analysis

A distinct set of risks emerges when computational tools – particularly natural language processing models and large language models – are applied to this historical corpus.

The most subtle is lexical semantic change: the gradual evolution of word meanings over time. Political language is especially susceptible to such shifts. Terms that carried specific connotations in the early twentieth century may have acquired different meanings in contemporary usage, while entirely new concepts have entered the political lexicon. A word like "liberal" meant something quite different in 1901 than it does today, particularly in Australia where it is now closely associated with the centre-right Liberal Party; "welfare" has shifted from a general term for wellbeing to a specific policy domain. When NLP models trained on modern corpora are applied to historical texts, they risk misinterpreting these shifting meanings in ways that are difficult to detect and even harder to correct. At the same time, these semantic shifts are not only a methodological hazard but also a substantive object of analysis in their own right: changes in the relational structure of political language can reveal how concepts are redefined, contested, and repositioned over time (Kozlowski et al., 2019).

Beyond vocabulary, there is the risk of misinterpreting the positions expressed in parliamentary speech. As the Background section established, parliamentary debate is rarely spontaneous deliberation – it is a choreographed performance shaped by party discipline, strategic communication, and awareness of the recorded audience. A speech may not represent the speaker's personal views, but rather the party line, a tactical intervention, or a performance designed for media consumption. Automated sentiment

analysis or stance detection applied naively to Hansard risks treating these strategic utterances as sincere expressions of belief.

These risks are compounded when large language models are used to summarise, classify, or interpret historical parliamentary text. Such models encode the linguistic patterns of their training data – predominantly contemporary text – and may impose modern framings on historical debates. They lack the contextual knowledge to recognise when a term has shifted meaning, when a speech is tactical rather than sincere, or when an apparent consensus masks deep disagreement conducted through other channels. The fluency of their outputs can mask these interpretive failures, producing confident summaries that are subtly but systematically wrong.

Conclusion

Hansard DB transforms over a century of Australian parliamentary proceedings into a structured, queryable resource that enables new forms of longitudinal analysis. By integrating biographical metadata with each parliamentary contribution, Hansard DB supports sophisticated longitudinal analyses of participation, representation, and rhetorical change across individuals, parties, and historical periods.

Yet the capacity to ask these questions does not guarantee the capacity to answer them well. The documented limitations here – missing days, incomplete attribution, and imperfect interjection parsing – represent known unknowns that can be quantified and accounted for. The interpretive risks are more insidious: the mediated nature of Hansard, the strategic character of parliamentary speech, and the dangers of applying modern computational tools to historical text.

Findings derived from Hansard DB must be interpreted with disciplined humility: confidence in the patterns the data reveals, paired with caution about the claims it can sustain. Hansard DB is a tool for inquiry, not a source of ground truth. Used carefully, it offers unprecedented access to the textual record of Australian democracy; used carelessly, it risks producing confident but misleading conclusions about what parliamentarians said, meant, and believed.

Data Availability

The full database, along with schema, metadata, usage instructions, and supporting code, is openly available at https://github.com/Fonzzy1/federal-hansard-db.

The version referenced in this paper is v4.1.0; however, it is expected that there will be frequent updates to include new parliamentary records.

Credit Statement

CRediT: Conceptualization: AC, SDA; Data curation: AC; Formal Analysis: AC; Funding acquisition: AC; Investigation: AC; Methodology: AC; Project administration: AC; Resources: AC; Software: AC; Supervision: SDA, LL; Validation: AC; Visualization: AC, SDA; Writing – original draft: AC; Writing – review & editing: AC, SDA, LL

References

AEC. (n.d.). Forming federal Government [[Accessed 10-02-2026]].

Alasuutari, P. (2025, February). The frontstage and backstage of parliamentary politics. In *National parliaments as a global institution* (pp. 108–124). Oxford University PressOxford. https://doi.org/10.1093/oso/9780192843623.003.0005

Alexander, R., & Alexander, M. (2021). The increased effect of elections and changing prime ministers on topics discussed in the australian federal parliament between 1901 and 2018. https://arxiv.org/abs/2111.09299

Angus, S. D. (2026). Tracking policy-relevant narratives of democratic resilience at scale: From experts and machines, to ai and the transformer revolution. *Data & Policy*, *8*. https://doi.org/10.1017/dap.2026.10063

Bäck, H., Baumann, M., Debus, M., & Müller, J. (2019). The unequal distribution of speaking time in parliamentary-party groups. *Legislative Studies Quarterly*, *44*(1), 163–193. https://doi.org/10.1111/lsq.12222

Bäck, H., & Debus, M. (2016). *Political parties, parliaments and legislative speech-making*. Palgrave Macmillan. https://doi.org/10.1057/9781137484550

Bayley, P. (2004). Introduction: The whys and wherefores of analysing parliamentary discourse. In P. Bayley (Ed.), *Cross-cultural perspectives on parliamentary discourse* (p. 1). John Benjamins Publishing Company.

Bowler, S., Farrell, D. M., & Katz, R. S. (1999). Party cohesion, party discipline, and parliaments. In *Party discipline and parliamentary government* (pp. 3–22). Ohio State University Press. Retrieved May 4, 2026, from http://www.jstor.org/stable/j.ctv177tghd.5

Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics*, *32*(10), 1439–1465.

Chisholm, E. (2005). Mps, the media, and the televising of parliament. *Political Science*, *57*(2), 65–73. https://doi.org/10.1177/003231870505700207

Coleman, J. (2023). Odd jobs: Parliament's hansard reporters 'learn about 20 different topics in a day' [Accessed: 2026-04-01].

Crewe, E. (2020, May). *The house of commons: An anthropology of mps at work*. Routledge. https://doi.org/10.4324/9781003086994

Davis, A. (2009a). Evaluating communication in the british parliamentary public sphere. *The British Journal of Politics and International Relations*, *11*(2), 280–297. https://doi.org/10.1111/j.1467-856x.2008.00344.x

Davis, A. (2009b). Journalist–source relations, mediated reflexivity and the politics of politics. *Journalism Studies*, *10*(2), 204–219. https://doi.org/10.1080/14616700802580540

Denniss, R. (2025). *Dead centre : How political pragmatism is killing us*. Australia Institute Press.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Edelman, M. J. (1988). *Constructing the political spectacle* [Published: 1988]. University of Chicago Press.

Edwards, C. (2016). The political consequences of hansard editorial policies: The case for greater transparency. *Australasian Parliamentary Review*, *31*(2), 145–160.

Feldman, C. (2023). You can't print that in hansard: Surveying hansard expungements in canada, australia and new zealand. *Australasian Parliamentary Review*, *38*(2), 97–117.

Hames, S., Haugh, M., & Musgrave, S. (2025). "how is that unparliamentary?": The metapragmatics of 'unparliamentary' language in the australian federal parliament. *Lingua*, *320*, 103932. https://doi.org/10.1016/j.lingua.2025.103932

Hebden, N., & Perche, D. (2023). Looking through the 'window on the house': Assessing the standard of question time in the australian house of representatives, 1991–2020. *Australian Journal of Political Science*, *58*(4), 343–362. https://doi.org/10.1080/10361146.2023.2224229

Ilie, C. (2022). How to argue with questions and answers: Argumentation strategies in parliamentary deliberation. *Languages (Basel)*, *7*(3), 205.

Karlsson, C., Persson, T., & Mårtensson, M. (2022). Do members of parliament express more opposition in the plenary than in the committee? comparing frontstage and backstage behaviour in five national parliaments. *Parliamentary Affairs*, *77*(1), 173–195. https://doi.org/10.1093/pa/gsac016

Karpowitz, C. F., & Mendelberg, T. (2014). *The silent sex: Gender, deliberation, and institutions*. Princeton University Press.

Katz, L., & Alexander, R. (2023). Digitization of the australian parliamentary debates, 1998–2022. *Scientific Data*, *10*(1), 567. https://doi.org/10.1038/s41597-023-02464-w

Katz, L., De Angelis, I., & Alexander, R. (2026). Do women politicians face more interruptions? an analysis of interjections in the australian parliamentary debates (1998-2025). *Australian Journal of Political Science*, 1–27. https://doi.org/10.1080/10361146.2026.2642819

Kotze, H., Korhonen, M., Smith, A., & van Rooy, B. (2023). Chapter 2. salient differences between australian oral parliamentary discourse and its official written records: A comparison of 'close' and 'distant' analysis methods. In M. Korhonen, H. Kotze, & J. Tyrkkö (Eds.), *Exploring language and society with big data: Parliamentary discourse across time and space* (pp. 54–88). John Benjamins Publishing Company. https://doi.org/10.1075/scl.111.02kot

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905â€“949. https://doi.org/10.1177/0003122419877135

Leslie, P. (2024). ausPH: An R package to retrieve data from the Parliamentary Handbook of the Commonwealth of Australia. [original-date: 2022-07-29T00:58:54Z]. https://github.com/palesl/ausPH

Martínez Arranz, A., Zech, S. T., & Bonotti, M. (2023). Political parties and civility in parliament: The case of australia from 1901 to 2020. *Parliamentary Affairs*, *77*(2), 371–399. https://doi.org/10.1093/pa/gsad008

Mclellan, R. (2019). Take off those olympic mittens, but the goldfish bowl is in order: Props, exhibits and displays in parliaments. *Canadian Parliamentary Review*, *42*, 11–16. https://api.semanticscholar.org/CorpusID:216716101

Miller, M. G., & Sutherland, J. L. (2023). The effect of gender on interruptions at congressional hearings. *American Political Science Review*, *117*(1), 103–121. https://doi.org/10.1017/S0003055422000260

Mollin, S. (2007). The hansard hazard: Gauging the accuracy of british parliamentary transcripts [Online]. *Corpora*, *2*(2), 187.

Nethercote, J. R. (2016, June). Australia's distinctive governance. In *Only in australia* (pp. 266–288). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198753254.003.0014

Ng, S., Zhang, J., Yu, S., Bhatti, A., Backholer, K., & Lim, C. P. (2025). Stance classification: A comparative study and use case on australian parliamentary debates. *Journal of Computational Social Science*, *8*(2), 43. https://doi.org/10.1007/s42001-025-00366-y

Parliament of Australia. (n.d.-a). Australian parliament Hansard [[Accessed 16-04-2026]].

Parliament of Australia. (n.d.-b). *The australian system of government* [Infosheet No. 20, House of Representatives Procedure Office]. Parliament of Australia. Canberra, Australia. https://www.aph.gov.au/-/media/05_About_Parliament/57_Education_Resources/571_Infosheets/PDF/is20.pdf

Parliament of Australia. (n.d.-c). *The house, government and opposition* [Infosheet No. 19, Procedure Office, House of Representatives]. Education Infosheet. Parliament

of Australia. Canberra, Australia. https://www.aph.gov.au/-/media/05_About_Parliament/57_Education_Resources/571_Infosheets/PDF/is19.pdf

Parliament of Australia. (n.d.-d). Question time [[Accessed 20-04-2026]].

Parliament of Australia. (n.d.-e). Quorum [Accessed: 2026-04-01].

Parliamentary Education Office. (2026). Standing orders [[Accessed 20-04-2026]].

Parliamentary Education Office. (n.d.). Hansard [Accessed: 2024-04-27].

Proksch, S.-O., & Slapin, J. B. (2014, December). *The politics of parliamentary debate: Parties, rebels and representation*. Cambridge University Press. https://doi.org/10.1017/cbo9781139680752

Sherratt, T. (2024). Commonwealth hansard xml repository updates. *Tim Sherratt*. https://updates.timsherratt.org/2024/05/26/commonwealth-hansard-xml.html

Slembrouck, S. (1992). The parliamentary hansard 'verbatim' report: The written construction of spoken discourse. *Language and Literature: International Journal of Stylistics*, *1*(2), 101–119. https://doi.org/10.1177/096394709200100202

Smith, R. (2018). The website "theyworkforyou" is changing british politics and not always for the better [Online article]. *TheyWorkForYou*. https://www.prospectmagazine.co.uk/politics/41257/the-website-theyworkforyou-is-changing-british-politics

Stiglitz, J. E. (1999). On liberty, the right to know, and public discourse: The role of transparency in public life. https://documents1.worldbank.org/curated/en/436941546609601734/pdf/WP-Stiglitz-right-to-know-OUO-9.pdf

Thompson, E. (2001). *UNIVERSITY OF NEW SOUTH WALES LAW JOURNAL*, *24*(3), 657–669. https://search.informit.org/doi/10.3316/informit.115821956554136

Tucker, E. C., Capps, C. J., & Shamir, L. (2020). A data science approach to 138 years of congressional speeches. *Heliyon*, *6*(8), e04417. https://doi.org/10.1016/j.heliyon.2020.e04417

Urbinati, N. (2006). *Representative democracy: Principles and genealogy* [See especially chapter on "The Diarchy of Will and Opinion"]. University of Chicago Press.

Vliegenthart, R., Walgrave, S., Baumgartner, F. R., Bevan, S., Breunig, C., Brouard, S., Bonafont, L. C., Grossman, E., Jennings, W., Mortensen, P. B., Palau, A. M., Sciarini, P., & Tresch, A. (2016). Do the media set the parliamentary agenda? a comparative study in seven countries. *European Journal of Political Research*, *55*(2), 283–301. https://doi.org/10.1111/1475-6765.12134

Ward, A. J. (2014). *Parliamentary government in australia*. Anthem Press.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502