Construction of a Language Resource Package of the Minutes of the National Diet of Japan for the Full-Text Search System "Himawari"

国立国語研究所論集 (NINJAL Research Papers) 22: 177–195 (2022)
ISSN 2186-1358

全文検索システム『ひまわり』用『国会会議録』パッケージの構築

山口昌也

国立国語研究所 研究系 音声言語研究領域

要旨

本稿は，『国会会議録検索システム』に収録されている国会会議録のテキストデータに基づき，全文検索システム『ひまわり』用の『国会会議録』パッケージを構築する方法，および，構築結果を報告する。本パッケージには，1947（第 1 回）～ 2012 年（第 182 回）に開催された衆議院・参議院の本会議，および，予算委員会の会議録 11106 件（約 4.49 億字）を収録している。本パッケージは言語表現の経年変化分析を行うために設計され，会議情報，発言者情報，会議録の構造情報が XML で付与されている。本稿では，まず，XML タグを設計するとともに，原資料の表記上の手がかりを使って，設計したタグを会議録に自動的にアノテーションする方法を示す。次に，考案した手法に基づいて『国会会議録』パッケージを構築する。また，構築したパッケージに収録した会議録の基礎情報を示す。最後に，『国会会議録』パッケージを使って，(a) 経年変化が大きい表現を抽出する方法，(b) 抽出された表現に対する経年変化要因を調査する方法を示すことにより，『国会会議録』パッケージの有用性を示す *。

キーワード：国会会議録，言語資料，全文検索システム『ひまわり』，経年変化分析

1. はじめに

筆者はこれまで国会会議録を研究資料として，言語表現の経年変化分析を行ってきた（山口 2017，山口 2019）。その過程で，『国会会議録検索システム』[1] に収録されている国会会議録のテキストデータを全文検索システム『ひまわり』[2]（山口・田中 2005）で利用できるよう整備を進めてきた。本稿では，『ひまわり』用の『国会会議録』パッケージの構築方法，および，構築結果を報告する。

国会会議録は，戦後から現在までの話し言葉を文字化した貴重な資料として，これまでに，多くの言語研究で利用されている（松田（編）2008，服部 2014 など）。経年変化の研究を行う上で国会会議録が優れている点は，対象とする議題や社会的環境の変化があるにしても，国会という単一の場において，規定された役割を持った話者が，規定された手順にしたがって行った議論を，長期間に渡って記録している点である。これは，言語表現の経年変化分析を行う上で，国会会議録を研究対象として選定した理由でもある。

* 本研究の一部は，国立国語研究所の共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）の研究成果である。また，本論文の一部は，山口 2017 で行った発表をもとにしている。『国会会議録』パッケージの作成にあたっては，原資料の利用の許可をくださった国立国会図書館，および，発言者情報を提供してくださるとともに，有益なアドバイスをくださった，同志社女子大学の故服部匡氏に心よりお礼を申し上げる。

[1] https://kokkai.ndl.go.jp/（2021 年 8 月 8 日確認）

[2] https://csd.ninjal.ac.jp/lrc/index.php?himawari（2021 年 11 月 22 日確認）

国会会議録のテキストデータを全文検索システム『ひまわり』用のパッケージとして構築するのは，言語研究時の検索の利便性とデータ利用の汎用性を考慮したためである。『ひまわり』は，言語研究用に開発された全文検索システムで，PC 用のデスクトップアプリケーションである。コンコーダンサとしての機能が主な機能であるが，単語検索，正規表現検索，検索結果の集計，資料に付与されている言語研究用の情報の集計，資料（国会会議録の場合は会議録）の Web ブラウザでの閲覧など，言語研究時の検索に役立つ機能が搭載されている。

また，「『ひまわり』用のパッケージ」としているが，後述するように，XML で記述されたテキストデータなので，汎用性が高く，テキスト処理の知識を持ったユーザであれば，『ひまわり』以外のツールでの利用も可能である。

『国会会議録』パッケージの開発は 2014 年頃から断続的に行っているが，開発前の段階で，すでに『国会会議録検索システム』を利用することができた。しかし，『国会会議録検索システム』の検索結果である会議録は，言語研究用に設計されていないため，例えば，次のような問題があった[3]。

- ある表現の二つの年の出現頻度を比較するために，年ごとの調整頻度を計算する際などには，『国会会議録検索システム』から年単位で会議録のデータをダウンロードする必要があった。
- 会議録が機械可読な形式で明示的に構造化されておらず，スクリプト言語などによる機械的なデータ処理には適していなかった。例えば，会議録には，参加者の発話だけでなく，出席者リストなどの会議関連情報が含まれているが，これらを区別して検索するのが困難だった。
- 年齢や肩書など発言者に関する情報が付与されていなかった。

そこで，『国会会議録検索システム』から取得した会議録のテキストデータを基に，経年変化分析で必要となる情報を XML により構造化された形式で記述し，『ひまわり』用の『国会会議録』パッケージとしてインターネット上に公開しつつ，改善を続けている。『国会会議録』パッケージに収録される会議録は，1947 年～ 2012 年までの衆議院・参議院の本会議，予算委員会の全会議録である。

『国会会議録』パッケージの特徴は，（1）本会議・委員会単位で 1947 年～ 2012 年の全会議録を収録することにより，経年変化分析を容易にしていること，（2）前述の構造化に際しては，発言部分を抽出しやすいよう，会議録の文書構造に対するアノテーションを行っている他，発言者名の正規化，生年・肩書の付与など，発言部分に対する経年的な分析に配慮した整備を行っていること，（3）前述のとおり，『ひまわり』用のパッケージとすることにより，構造化された情報を検索時や検索結果の集計時に利用できることである。

本稿の構成は，次のとおりである。まず，2 節では『国会会議録』パッケージの設計として，パッケージの構成，収録データのほか，XML タグの仕様を決定する。次に，3 節では実際にパッケージを構築する方法として，原資料である会議録の収集方法，設計したタグを自動的にアノテーショ

[3] 現在は，「国会会議録検索システム　検索用 API」（川瀬・清水 2015，国立国会図書館 2019）が公開されているため，一部の問題は解決されている。

ンする方法，発言者の生年情報を付与する方法などについて説明する。4 節では，構築結果の『国会会議録』パッケージの概要を示し，続く 5 節では，パッケージの有用性を示すために，パッケージを用いた，（a）経年変化の大きな表現の抽出方法，（b）抽出した表現に関する経年変化要因を調査する方法を示す。そして，最後に，6 節で本稿のまとめを述べる。

2. 『国会会議録』パッケージの設計

2.1 『国会会議録』パッケージの構成と収録データ

前述のとおり，『国会会議録』パッケージには，衆議院・参議院両院の本会議，予算委員会の計四つの会議録を収録している。収録期間は，第 1 回国会が開催された 1947 年から第 182 回国会が開催された 2012 年までである。原資料は，『国会会議録検索システム』において号単位で公開されていた会議録の HTML ファイルである。なお，収録にあたっては，『国会会議録検索システム』を運営している国立国会図書館から，言語資料としての整備と，整備したデータを再配布するための許諾を得ている。

多数ある会議のうち，本会議と予算委員会を選択した理由は，言語の経年変化の分析という目的のもと，（a）継続性，（b）重要性，（c）対話スタイルの多様性を考慮したことによる。まず，（a）の継続性については，本会議，予算委員会ともに第 1 回から存続する会議であることから選択した。（b）の重要性については，本会議が全議員による会議であり，それぞれの院の最終的な議決を行う場である点，予算委員会は（a）の条件を満たす常任委員会である点を考慮した。（c）については，本会議と予算委員会との対話スタイルの違いに着目した。具体的には，全議員参加の本会議では，国務大臣の演説や委員会報告といったような，参加者全員に向けた発言が含まれる会議なのに対して，予算委員会では，基本的に特定の聞き手を想定した質疑・応答が主体の会議である。

『国会会議録』パッケージ全体の構成としては，表 1 のように，院（衆議院・参議院）と会議（本会議・予算委）の組み合わせで，大きく四つに分割した。院と会議の別だけを考慮するならば，この四つをサブコーパスとするのが一般的であると思われるが，（a）第 1 回～第 144 回までは，画像から機械的に文字を読み取って作成されていること，（b）動作環境によっては『ひまわり』で処理できないサイズのコーパスになってしまう [4] ことから，さらに開催回により，（i）第 1 回～第 48 回，（ii）第 49 回～第 96 回，（iii）第 97 回～第 144 回，（iv）第 145 回～第 182 回の四つに分割し，計 16 個のサブコーパスから構成することとにした。なお，『ひまわり』では複数のサブコーパスを選択して，一括で検索などを行うことができるため，利用上，特に問題は発生しない。

[4] 特に，使用可能メモリが少ない場合や，32 ビット版の OS の場合に，メモリ不足でエラーが発生する。ただし，『ひまわり』の検索処理はサブコーパスごとに検索処理を行い，結果を統合するので，サブコーパスに細かく分割すれば，メモリ不足の問題は軽減される。

表 1 『国会会議録』パッケージの構成

	本会議	予算委員会
衆議院	第 1 回～第 48 回　第 49 回～第 96 回 第 97 回～第 144 回　第 145 回～第 182 回	第 1 回～第 48 回　第 49 回～第 96 回 第 97 回～第 144 回　第 145 回～第 182 回
参議院	第 1 回～第 48 回　第 49 回～第 96 回 第 97 回～第 144 回　第 145 回～第 182 回	第 1 回～第 48 回　第 49 回～第 96 回 第 97 回～第 144 回　第 145 回～第 182 回

2.2 会議録の構造

本節では，アノテーションの設計を行う前に，原資料である会議録の文書構造を示しておくことにする。例として，第 34 回国会・衆議院・予算委員会（第 7 号，1960 年 2 月 11 日）の会議録全体の構造を図 1 に示す。

図 1 に示したとおり，会議録の構造は，大きく分けて，次の三つに分かれている。この構造は，両院の本会議，予算委員会にかかわらず，おおむね共通している [5]。

- ヘッダ部：発言部の前に，会議の名称，開催日時，会議の参加者，議題など，会議に関する情報が記述されている。なお，会議の開催に至らなかった場合（例：第 71 回国会・衆議院・本会議・第 64 号，上から 9 番目の徳永正利氏による発言部分）や本会議の開会式の場合（第 138 回国会・衆議院・本会議・第 3 号）などは，ヘッダ部しかない場合もある。
- 発言部:会議における参加者の発言が記述されている。ただし，発言者本人の発言だけでなく，関連資料（例：第 98 回国会・参議院・本会議・第 14 号），拍手などの状況描写，「異議なし」やヤジといった発言者以外の発言などの付属情報が含まれる場合がある。例えば，図 1 の例では，「〔「異議なし」と呼ぶ者あり〕」という付属情報を含んでいる。
- フッタ部：発言部のあとに，会議の関連情報が追記される場合がある。図 1 末尾のように，会議の散会時刻が記述されていることが多いが，出席者一覧（例：第 10 回国会・参議院・予算委員会・第 33 号），付帯決議（例：第 1 回国会・衆議院・予算委員会・第 6 号），開会式の情報（例：第 119 回国会・衆議院・本会議・第 1 号）などが記述されている場合もある。

発話の分析を行う場合は，発言部のみが対象になると思われるが，会議録としての一貫性を考慮し，各部が明確に区別できるようアノテーションした上で，『国会会議録パッケージ』にはすべてを収録することにする。

[5] ただし，内部の記述内容や構造には違いがある。例えば，予算委員会にはヘッダ部に参加者が記載されているが，本会議では記載されていない。また，衆議院・本会議ではフッタ部に出席国務大臣が列挙される場合がある。

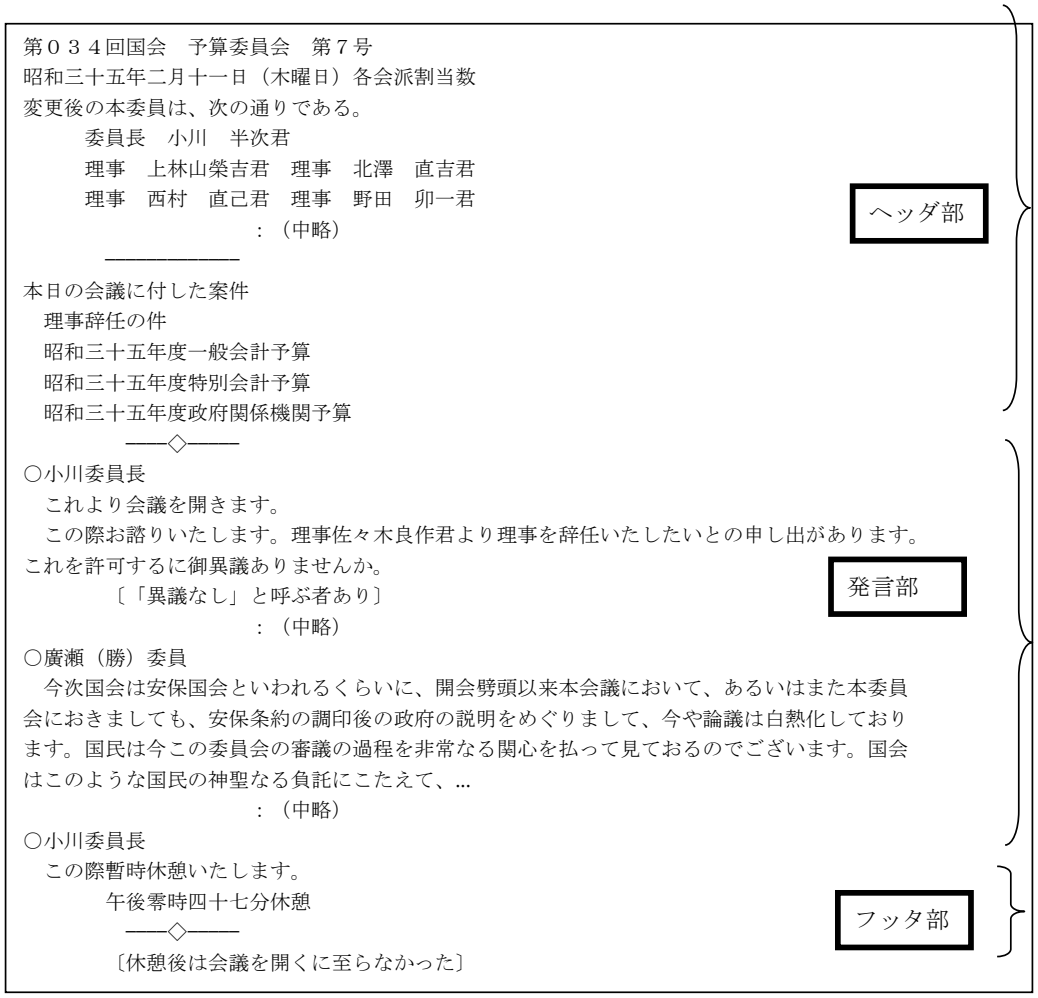

図 1　会議録の例（第 34 回国会・衆議院・予算委員会）

2.3　付与する研究用情報

『国会会議録』パッケージは経年変化分析での利用を想定している。そのため，研究用情報は，(a) 会議録から発言部分をそれ以外の部分と区別して検索できること，(b) 発言の話者，および，発言した会議の情報を取得できることに配慮して設計した。研究用情報の記述は，XML で記述する。使用したタグの一覧を表 2 に示す。なお，表中の属性は，属性名（説明）という形式で表記している。例えば，corpus タグの「name（コーパス名）」は，name 属性がコーパス名を表すことを示す。

表 2 『国会会議録』パッケージで使用したタグ一覧

タグ名	タグの概要と属性
corpus	概要：一つのサブコーパスに相当する。 属性：name（コーパス名），ver（バージョン番号）
minutes	概要：一つの会議録に相当し，header, body, footer 要素からなる。 属性：title（会議録のタイトル），jdate（開催日・原資料での表記），house（衆議院・参議院の別），no（開催回），meeting（本会議・予算委員会の別），vol（開催号），date（開催日・year-month-day 表記），url（原資料の URL），nchar_all（会議録の総文字数），nchar_body（発言部の総文字数）
header	概要：会議録のヘッダ部，属性：なし
body	概要：会議録の発言部，属性：なし
footer	概要：会議録のフッタ部，属性：なし
utterance	概要：発言部に含まれる 1 回の発言 属性：speaker_org（発言者名・原資料での表記），speaker（発言者名・漢字正規化前），speaker_norm（発言者名・漢字正規化後），birth_year（西暦での生年），title（肩書）
1	概要：header，utterance, footer 会議録における行，属性：なし
info	概要：utterance 要素に含まれる発言以外の要素（非発話要素） 属性：text（原資料での当該文字列）

『国会会議録』パッケージ中のサブコーパスは corpus 要素[6] をルートノードとする単一の XML ファイルとして記述される。サブコーパスは図 2 のように，収録対象の会議録（minutes 要素）の集合として構成する。会議録の内部構造は，図 1 で示した構造をそのまま反映し，ヘッダ部（header 要素），発言部（body 要素），フッタ部（footer 要素）からなり，発言部は発話の集合で構成される。

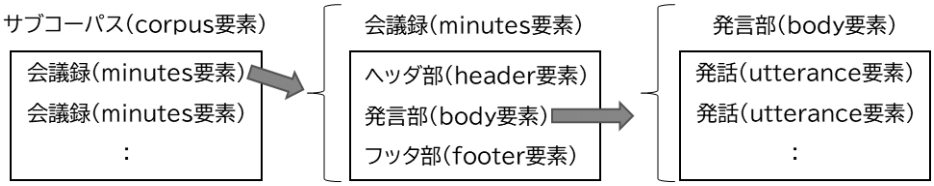

図 2 サブコーパスの内部構造

例として, 図 1 の会議録を表 2 のタグでアノテーションした結果を図 3 に示す[7]。冒頭の minutes タグでは，属性で会議種別や開催日などの情報を記している。minutes 要素内部では，先頭から順にヘッダ部，発言部，フッタ部をそれぞれ header, body, footer タグでマークアップする。

header 要素と footer 要素中の各行は 1 タグでマークアップしている。検索のみの利用であれば必要ないタグであるが，会議録全体を表示する際などに，行の情報が必要になるため，付与している。

body 要素内の発言は utterance タグでマークアップしている。この際，タグの属性として，発

[6] タグ X でマークアップされる要素を「X 要素」と表記する。
[7] 説明に関係ない記述は一部省略。

```
<minutes title="衆議院会議録情報　第０３４回国会　予算委員会　第７号" jdate="昭和三十五年二月十一日"
house="衆議院" no="034" meeting="予算委員会" vol="07" date="1960-02-11"
 url="http://..." nchar_all="33782" nchar_body="31962">
<header>
<l>第０３４回国会　予算委員会　第７号</l>
<l>昭和三十五年二月十一日（木曜日）各会派割当数</l>
<l>変更後の本委員は、次の通りである。</l>
              :
</header>
<body>
<utterance speaker_org="小川委員長" speaker="小川半次" speaker_norm="小川半次"
 birth_year="1909" title="委員長">
<l>　これより会議を開きます。</l>
<l>　この際お諮りいたします。理事佐々木良作君より理事を辞任いたしたいとの申し出があります。これ
を許可するに御異議ありませんか。</l>
<l><info text="　　〔「異議なし」と呼ぶ者あり〕" /></l>
              :
</utterance>
</body>
<footer>
<l>　　午後零時四十七分休憩</l>
<l>　　――◇――</l>
<l>　　〔休憩後は会議を開くに至らなかった〕</l>
</footer>
</minutes>
```

図 3　アノテーション例

言者名，発言者の生年，肩書が付与される。発言者名については，検索の便を考慮し，会議録上の発言者表記（例：小川委員長，speaker_org 属性に記述），会議録上の氏名表記（例：小川半次，speaker 属性に記述），正規化された氏名表記（speaker_norm 属性に記述）の 3 種類の情報が併記される。氏名の正規化は，常用漢字変換を用いた（例：「佐藤榮作」→「佐藤栄作」）。また，発言者名は本文[8]からは除外し，属性として記述していることにも注意されたい。これは，全文検

[8] XML のテキスト要素。

索時に発言者名と本文を区別して検索できるよう，配慮したためである。なお，発言が複数行にわたる場合は，header, footer 要素と同様 1 タグで行の情報を付与している。

info タグは，utterance 要素中の発言以外の要素をマークアップするために用いる空要素タグである。図 3 では，「〔「異議なし」と呼ぶ者あり〕」の部分で用いられている。info タグの text 属性値として記述することにより，非発話要素として，本文から除外する。これにより，『ひまわり』の全文検索の対象とはならなくなる。

3.『国会会議録』パッケージの構築

3.1 概要

本節では，『国会会議録検索システム』から取得した会議録データから『ひまわり』用の『国会会議録』パッケージを構築する方法について説明する。おおまかな構築の流れは，（1）原資料の収集，（2）自動アノテーション，（3）話者生年情報の付与，（4）パッケージ化，である。後続の節で，個々に詳しい内容を説明する。

なお，『国会会議録』パッケージ構築を最初に公開した後の 2014 年 12 月より「国会会議録検索システム　検索用 API」（川瀬・清水 2015，国立国会図書館 2019）が公開されている。そのため，新規に構築する場合は，異なった構築方法になる可能性がある。

3.2 原資料の収集

原資料の収集には，ダウンローダ[9]を使用し，『国会会議録検索システム』サイトから HTML ファイルをダウンロードした。原資料の収集は，2014 年 3 月 27 日から 28 日の 2 日の間に実施した。原資料の一覧は，『国会会議録検索システム』の「選択閲覧」[10] のリンクに基づいて取得した。収集した会議数は，合計 11106 である。内訳については，4.1 節を参照されたい。

3.3 自動アノテーション

前述のとおり，原資料の HTML ファイルには，表 2 のタグに相当する情報が明示的に記述されていないため，原資料である会議録の表記上の手がかりをもとにスクリプト言語[11]で必要な情報を抽出することにより，自動的にアノテーションした。以下の節では，その詳細を示す。

3.3.1 マークアップの範囲

corpus, minutes, l タグは，それぞれサブコーパス，会議録，会議録中の行にマークアップを行うため，その範囲は明確である。一方，header, body, footer, utterance, info タグは，会議録の表記上のルールに基づいて，マークアップの範囲を決める必要がある。表記上のルールは，明文化さ

[9] wget を使用した。https://www.gnu.org/software/wget/（2021 年 8 月 8 日確認）
[10] 本稿執筆時点（2021 年 6 月 10 日）ではページ自体がすでに存在しないが，URL は http://kokkai.ndl.go.jp/SENTAKU/index.htm である。Internet Archive（https://web.archive.org）などから参照されたい。
[11] プログラミング言語 Perl を使用した。

れておらず，院，会議種別によって異なるだけでなく，開催時期によって変わってくる場合もあるため，テストと微調整を繰り返しつつ，変換ルールを作成した。使用した変換ルールの一部を次に示す。ただし，これらのルールは経験則であり，誤った変換をする可能性もある。

- 発言の冒頭は，「<B> ○小川委員長 </B>」のように，HTML の B タグでマークアップされ，「○」＋発話者名の形式で記述されている行から次の発言，もしくは，会議末までを utterance 要素として，マークアップする。
- header 要素は，原資料の先頭から，最初の発言の前までをマークアップする。
- 最後尾の utterance 要素の中に，次のような条件を満たす文字列があった場合は，当該 utterance 要素の範囲をその直前に修正して，それ以降を footer 要素とする。
  - ➢「　　午後零時四十七分休憩」や「　　――――――――――――――――――――」のように行頭が空白文字で始まり，「午前」「午後」「――」が続く行
- 最初の utterance 要素から footer 要素の直前までを body 要素とする。
- 次の条件を満たす文字列があった場合，本文ではなく，info 要素の text 属性値（非発話要素）とする。
  - ➢ utterance 要素中の行のうち，行頭が空白文字で始まり，「午前」「午後」「――」が続く行，もしくは，行頭が空白文字で始まり，括弧で囲われた文字列で終わる行（例えば，図 3 の〔「異議なし」と呼ぶ者あり〕の行全体）。なお，「――」が続く行の場合，次の行に句点を含まないなど，一定の条件を満たせば，当該の utterance 要素の残りの行すべてを info 要素とする。例えば，法律案や投票結果の名簿などがこの条件を満たすことが多い。
  - ➢ utterance 要素中の括弧で囲われた文字列で，「叫ぶ」「呼ぶ」「拍手」「騒然」「聴取不能」のような文字列を含むもの（例：「～であります。（拍手）」）。これらは，会議の様子を描写する記述の場合が多い。

3.3.2 発言者名，肩書

3.3.1 節で述べたように，原資料で発言の冒頭に記載されている発言者の情報は，本文からは除外し，utterance タグの speaker_org, speaker, speaker_norm 属性，title 属性として記述する。ただし，発言者名と肩書の表記方法が，衆議院・予算委員会とそれ以外で，次のように異なるため，個別の変換処理が必要となる。

- 衆議院・予算委員会の場合，「麻生国務大臣」「生方委員」のように「名字＋肩書」の形式で表記される。
- それ以外の会議録の場合，委員長や国務大臣などの肩書があれば，「委員長（平島敏夫君）」「国務大臣（安井謙君）」のように，「肩書＋（氏名＋「君」）」の形式で表記される。また，単に委員であれば，「羽田孜君」のように「氏名＋「君」」の形式で表記される。

このうち，後者の場合は，肩書があれば，氏名と肩書が明記されており，肩書が付与されていなければ委員なので，容易に utterance タグの speaker_org, speaker, title 属性値を取得できる。例えば，「国務大臣（佐藤榮作）」を取得できれば，それが speaker_org 属性値となり，そこから speaker 属性値の「佐藤榮作」，title 属性値の「国務大臣」を取り出す。さらに，speaker 属性値に対して常用漢字変換を施すことにより，speaker_norm 属性値の「佐藤栄作」を得る。

一方，衆議院・予算委員会の場合，肩書は全発言者についているものの，氏名のうち名前が省略されている。そのため，会議録ヘッダ部の会議情報から，名字の情報を抽出している。例えば，図 1 では，発言部分には「小川委員長」となっており，名前の「半次」が欠落している。一方，ヘッダ部をみると，「委員長　小川　半次君」とあるので，肩書の「委員長」と姓の「小川」を手がかりとして，名前の「半次」を抽出する。

3.3.3 会議情報

minutes タグの属性として記述する会議情報は，次のように取得している。

- 開催日（jdate 属性）は，原資料のヘッダ部で最初に現れた日付表示を抽出し，year-month-day 表記（例：1960-02-11）に変換した上で，date 属性値とした。
- 会議録のタイトル（title 属性）は，原資料の HTML ファイルの title 要素から取得し，そこから院（house 属性），回（no 属性），号（vol 属性），会議種別（meeting 属性）を抽出した。
- nchar_all，nchar_body 属性値は，ぞれぞれ minutes 要素に含まれる文字数，body 要素に含まれる文字数を計測した値である。

3.4 話者生年情報の付与

発言は utterance タグでマークアップし，birth_year 属性で話者の生年を記述する。ただし，話者の生年情報は会議録に記載されていないため，『国会会議録』パッケージでは，同志社女子大学の服部匡氏が作成し，氏のホームページ [12] で公開されていたデータを利用して，機械的に付与した。このデータは，参議院議員の生年・在職期間・本籍・出身地データを「参議院のウェブページ・『議会制度百年史』・『歴代国会議員名鑑』・『参議院要覧』・各種ウェブサイトによって調査した」ものである。このデータから生年情報を得られない発言者のうち，発言数 100 回以上の発言者に関しては，Wikipedia[13]，および，コトバンク（20 世紀日本人名事典）[14] などを調査して，生年を付与した。

付与結果を表 3 に示す。なお，ここで言う「発言者数」とは，utterance タグの speaker 属性

[12] http://thattori.com/（2021 年 6 月 10 日確認）。なお，2021 年 8 月 8 日確認時にはサイトが廃止されていたため，Internet Archive の Wayback Machine の URL も付記しておく。http://web.archive.org/web/20180825211919/http://thattori.com/asakura/asakura.html
[13] https://ja.wikipedia.org/（2021 年 8 月 8 日確認。ただし，構築に利用したのは主に 2016 年〜 2017 年）
[14] https://kotobank.jp/（2021 年 8 月 8 日確認。ただし，構築に利用したのは主に 2016 年〜 2017 年）

値[15]の異なり数である。発言者数で見ると，参議院・予算委員会で 35.5% と付与率が低いが，これは官僚や参考人など議員以外の発言者が存在することによる。そのため，議員の発言がほとんどを占める本会議では，付与率は参議院で 90.6%，衆議院でも 82.6% となる。また，発言数で見れば，最も低い参議院・予算委員会でも 89.5% の発言に発言者の生年情報が付与されていることになる。

表 3　発言者の生年付与結果

	発言者数（異なり）	生年付与数	発言数	生年付与数
参議院・予算委員会	5182	1838（35.5%）	483989	433401（89.5%）
衆議院・予算委員会	4525	1678（37.1%）	429153	389946（90.9%）
参議院・本会議	2066	1871（90.6%）	105638	105025（99.4%）
衆議院・本会議	2591	2140（82.6%）	119740	116955（97.7%）

3.5　全文検索システム『ひまわり』用のパッケージ化

『ひまわり』用パッケージファイルは，コーパスを格納した XML ファイルと検索用の索引ファイル，パッケージの設定ファイルなどを一定のフォルダ構造で格納した ZIP ファイルである。『国会会議録』パッケージでも，前節までに述べた方法で作成したサブコーパスの XML ファイルに対して，索引付けを行い，パッケージ化した。この際，利便性とダウンロード時間などを考慮し，本会議，予算委員会の二つに分けてパッケージ化し，『国会会議録』パッケージのホームページ[16]で公開している。

また，『ひまわり』には検索対象のコーパスを形態素解析し，解析結果をアノテーションする機能がある。『国会会議録』パッケージにもこの機能を適用し，「形態素解析結果の追加パッケージ」として公開している。形態素解析は外部のシステムを利用するようになっているため，目的に応じて，形態素解析をやり直すことも可能である[17]。なお，形態素解析結果のアノテーションは，コーパスを格納した XML ファイルに対して直接行うのではなく，別ファイルに独自形式のバイナリデータとして記録される。

【『国会会議録』パッケージを『ひまわり』で利用するには，パッケージの ZIP ファイルを『ひまわり』へドラッグ＆ドロップして，インストールする。】図 4 は，発言部に対して「してまいりましたが」を全文検索した結果である。ウィンドウ左側から検索文字列「してまいりましたが」の前後文脈のほか，検索文字列を含む会議録名，発言者名など，minutes，utterance 要素に付与した属性が表示されていることがわかる。また，検索結果をダブルクリックすることにより，図 4 のように，当該用例を含む会議全体を Web ブラウザで表示することができる。

[15] 服部匡氏のデータには，氏名に関して，異表記の情報も付与されているため，発言者名の照合には，正規化後の speaker_norm 属性ではなく，speaker 属性を用いた。そのため，表 3 は speaker 属性値の異なりで集計した。
[16] https://csd.ninjal.ac.jp/lrc/index.php?kokkai（2021 年 11 月 22 日確認）
[17] 公開している追加パッケージでは，形態素解析システム MeCab（ver.0.996，辞書は IPA）を使用した。https://taku910.github.io/mecab/（2021 年 8 月 8 日確認）

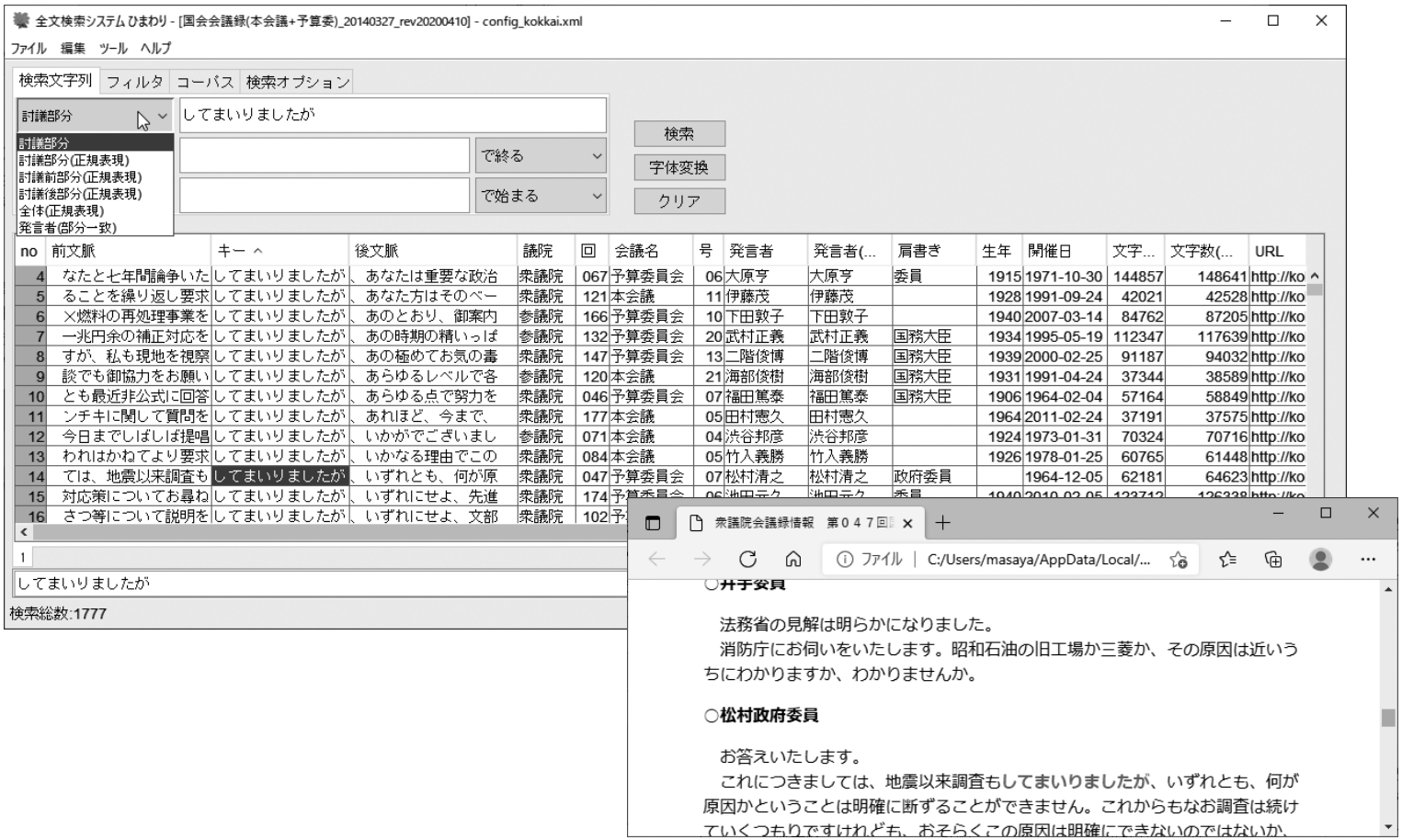

図 4 『ひまわり』での『国会会議録』パッケージの検索例

【検索方法については】図 4 左上のプルダウンメニューに表示されているとおり，ヘッダ部，フッタ部，発言部別[18]の検索や会議録全体の検索，発言者名での検索ができるようになっている。この設定は，パッケージ化の際に行っており，変更することもできる。例えば，info タグにより本文から除外された非発話要素を検索対象とすることも可能である。

4. 構築結果

4.1 会議録に関する基本的な情報

構築した『国会会議録』パッケージを概観するために，収録されている会議録に関する情報を表 4 に示す。

まず，『国会会議録』パッケージに収録した会議数は，表 4 のとおり，本会議，予算委員会がそれぞれ 7127，3979，合計 11106 である。文字数はパッケージ全体で約 4.49 億文字，発言部分に限定すると，約 4.34 億文字であった。発言数については，表 3 から，本会議が約 22.5 万回，予算委員会は約 91.3 万回である。この結果と文字数の結果より，1 回あたりの平均発言文字長は，本会議，予算委員会，それぞれ 633.3，319.0 文字となり，本会議のほうが約 2.0 倍長いことがわかる。

表 4 『国会会議録』パッケージに含まれる会議数，文字数，info タグ数

	会議数	文字数（発言部）	文字数（全体）	info タグ数	info タグ text 属性
参議院・予算委員会	1945	136357129	141307828	32765	445372
衆議院・予算委員会	2034	154889998	159987430	32974	566419
小計	3979	291247127	301295258	65739	1011791
参議院・本会議	3090	69117051	71432594	177340	2492660
衆議院・本会議	4037	73603580	76252552	308655	4703860
小計	7127	142720631	147685146	485995	7196520
総計	11106	433967758	448980404	551734	8208311

次に，発言部分に含まれる非発話要素の量を示す。これには，info タグの text 属性値を用いる。表 4 の「info タグ数」欄は info タグの付与数，「info タグ text 属性」欄は非発話要素として認定されることにより，本文から除外され，info タグの text 属性に記述された文字列長の合計である。この結果から，原資料の発言部のうち，本会議では約 4.8%，予算委員会では約 0.3% が非発話文字列だったことがわかる。網羅的な調査は行っていないが，本会議で非発話文字列が多いのは，「（拍手）」などの会議の状況描写としての非発話要素だけでなく，決議案や投票結果の名簿のような会議資料が予算委員会よりも多く含まれるためだと思われる。

さらに，開催年別の発言文字数を図 5 に示した。開催年別の平均文字数は，衆議院・参議院の本会議がそれぞれ約 111.5 万字，104.7 万字，予算委員会は衆議院･参議院それぞれ約 234.7 万字，

[18] 図 4 のプルダウンメニュー中の「討論前部分」「討論部分」「討論後部分」は，それぞれ「ヘッダ部」「発言部」「フッタ部」に対応する。

約 206.6 万字であった。

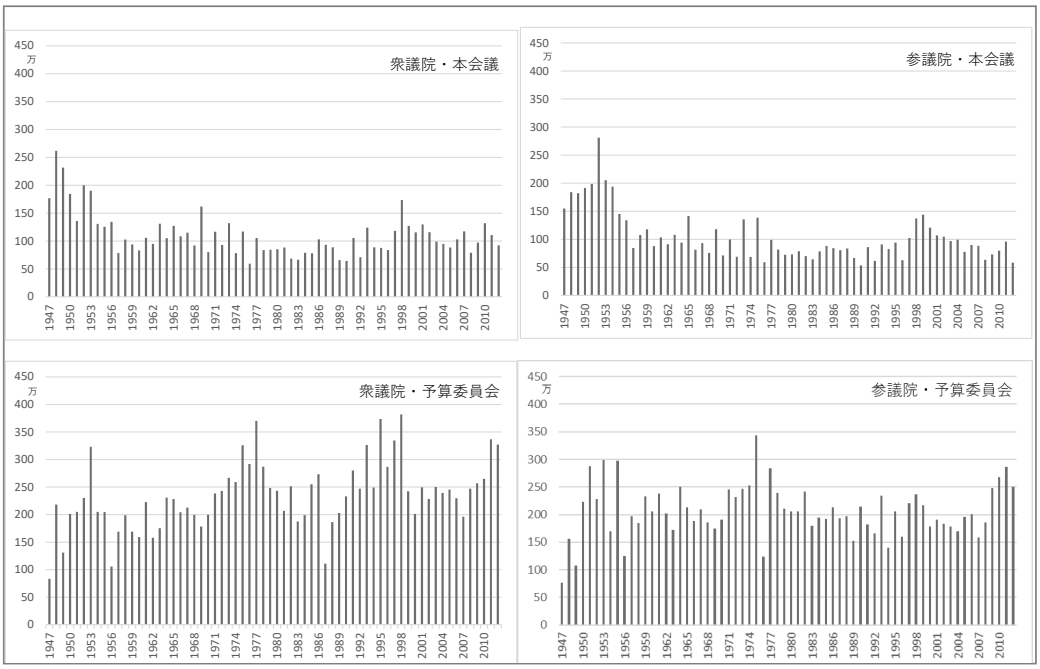

図 5　開催年別の発言文字数

4.2　発言者に関する基本的な情報

院，会議種別の発言者数（異なり）を発言者名の正規化前後で集計し，表 5 に示す。『国会会議録』パッケージ全体でみると，発言者名正規化後の発言者数の異なりは 7813 名であった。なお，同姓同名のチェックは行っていない。

表 5　発言者数（異なり）

	予算委員会		本会議	
	正規化前	正規化後	正規化前	正規化後
衆議院	4525	4469	2591	2567
参議院	5182	5122	2066	2053

次に，衆議院に関して，発言者の肩書（発言数上位 10 位まで）を，発言数・発言文字数と併記して，表 6 に示す。なお，本会議中の肩書［無表記］は，肩書の表記がなかったものであるが，ほとんど議員の発言であると思われる。また，肩書の表記については，正規化をしていないため，「国務大臣」と「國務大臣」のように，区別して集計されていることに注意されたい。

予算委員会と本会議を比較すると，(a) 本会議では，発言文字数は少ないものの，「議長」「副議長」の発言数が多い，(b) 発言文字数で見ると，予算委員会，本会議ともに，委員，国務大臣，

内閣総理大臣が主要な発言者であるが，予算委員会では，発言文字数が内閣総理大臣と同程度の政府委員をはじめとして，参考人，政府参考人，証人の発言文字数の割合が本会議よりも高いことがわかる。

表 6　肩書別の発言数と発言文字数（衆議院）

予算委員会			本会議		
肩書	発言数	発言文字数	肩書	発言数	発言文字数
委員	198696	90511105	議長	63046	3521690
国務大臣	111735	35681791	［無表記］	26819	48623277
内閣総理大臣	32149	10915837	副議長	13982	619973
委員長	31795	1897103	国務大臣	10528	11400633
政府委員	29785	9350421	内閣総理大臣	3395	8262481
参考人	6556	2340948	國務大臣	992	883194
証人	4998	600656	事務総長	367	33281
國務大臣	3481	1298557	政府委員	261	162137
政府参考人	3167	669460	仮議長	209	11893
委員長代理	2552	100679	参議院議員	32	55423

5. 『国会会議録』パッケージの使用例

5.1 頻度に関して経年変化の大きな表現の抽出

『国会会議録』パッケージの使用例として，頻度に関して経年変化の大きな表現を抽出してみる。抽出は，『国会会議録』パッケージに収録されている会議録のうち，収録期間の先頭と末尾の一定期間の資料に対して，文字 5gram を求め，頻度変化の大きな文字列を求めることにより行う。今回は，対象とする資料を衆議院･予算委員会，先頭と末尾の期間は第 1 回～ 48 回（1947 ～ 1965），第 145 回～ 182 回（1999–2012）とした。

文字列 s に対する変化率 r を次のように定義する。

$$r = (f_{tail} - f_{head}) / \max(f_{head}, f_{tail}) \times 100$$

なお，r は $-100 \leq r \leq 100$ の値を取り，$r > 0$ の場合は 2 期間の間で増加，$r < 0$ の場合は減少していることを表す。f_{head}，f_{tail} は，それぞれ，期間の前部・後部における 10 万字あたりの出現頻度である。

表 7，表 8 は，それぞれ，増加分の比率が大きい文字列，減少分の比率が大きい文字列である。ただし，低頻度の文字列は変動が大きくなる場合があるため，10 万字あたりの出現頻度が 20 以上の文字列に限定している。また, スペースの関係上,「おるのであ」と「ておるので」のように，重複を含む文字列は，基本的に変化率の絶対値の大きなほうしか表に挙げていない。

以上の抽出過程において，特に強調したいのは，f_{head}，f_{tail} を得る段階までは，『ひまわり』の標準機能のみで計算できる点である。具体的には，文字 5gram を作成するには，正規表現「.....」により，衆議院・予算委員会のデータから，長さ 5 文字のすべての部分文字列を検索し，検索結

果の「集計機能」[19] で個々の部分文字列の頻度を集計している。f_{head}, f_{tail} が得られれば，Microsoft Excel などの表計算ソフトウェアで容易に r を求めることができる。

表 7　増加分比率が高い文字列

文字列	f_{head}	f_{tail}	r
んですけれ	0.24	24.8	99
ふうに思い	0.24	21.4	98.8
思うんです	0.42	33.4	98.7
いるんです	0.67	39.8	98.3
ないんです	0.53	28.5	98.1
させていた	1.36	42.9	96.8
ています。	1.31	34.2	96.2
。そして、	0.94	22.9	95.9
そういった	1.69	22.9	92.6
てください	1.88	25.3	92.6
ころでござ	1.86	22.9	91.9
でしょうか	2.89	31.3	90.8
うに思いま	2.59	25.8	90
思っており	8.57	61.5	86.1
いるわけで	7.06	50.6	86.1
いました。	3.87	24.4	84.1
おっしゃっ	4.25	25.2	83.1
それから、	3.84	22.5	83
というのは	17.6	88.3	80.1
ないですか	5.39	23.8	77.3
されている	6.31	27.5	77.1
言っている	4.65	20.1	76.8
なっている	6.67	28.3	76.5
ですから、	14	53.7	74

表 8　減少分比率が高い文字列

文字列	f_{head}	f_{tail}	r
になつてお	22.9	0	-100
われわれは	21.8	0.02	-99.9
おるのであ	53.4	0.2	-99.6
、そうして	20.3	0.3	-98.5
ればならぬ	41.7	1	-97.6
おるという	20.3	3.2	-84.4
いたしたい	25.5	4	-84.4
のでござい	43	8.6	-80
いうものは	42.2	9	-78.7
おきまして	59.9	13.4	-77.7
思うのです	23.1	5.5	-76.2
しましては	26.1	7	-73.3
であります	472.9	126.9	-73.2
。こういう	26.5	7.7	-70.1

5.2　表現の経年変化の要因の調査

前節で経年変化の大きな表現を抽出した。次の段階として，変化の要因を調査するため，得られた表現の開催年別の出現頻度を計測する。ここでは，特徴的な変化が発生している例として，「になつてお」（表 8 の第 1 位）を題材に，促音の表記に関する変化を示すことにする。

図 6 は，「になつてお」「になってお」の開催年別の 10 万字あたりの調整頻度である。これを見ればわかるとおり，1955 年を境に，大きい「つ」を含む「になつてお」が使用されなくなり，「になってお」が出現するようになる。それに対して，小さい「っ」を含む候補，例えば，「そういった」（表 7 の第 9 位）では，1955 年以前はほとんど出現せず，「そういつた」が用いられている（図 7）。これにより，1955 年周辺で促音に関する表記規則になんらかの統一的な変更が行われたことが推測できる。

[19]『ひまわり』利用者マニュアル「4.4 検索オプション」・「抽出オプション」・「頻度計測のみ」・「一覧」を参照のこと。

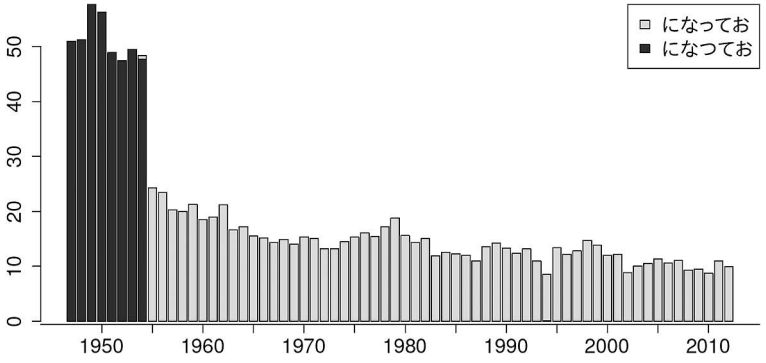

図 6 「になつてお」

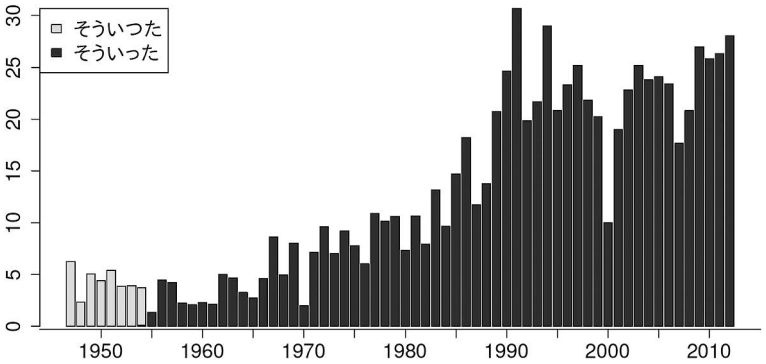

図 7 「そういつた」

したがって，抽出された文字列のうち，「思っており」「おっしゃった」「言っている」「なっている」なども同様に促音に関する表記規則の変更が変化要因になっている可能性が高くなる。紙面の関係上，例を示すだけにとどめるが，促音以外にも表記規則の変更が変化の要因になっている例として，「われわれは」と「我々は」（ひらがな表記と漢字表記），「思うのです」と「思うんです」（「の」と「ん」），「、そうして」と「。そして、」（ウ列の長音の有無）にも見られた。

筆者は経年変化のモデルを構築しようと試みているが，ここで示した方法を用いて，表記規則の変更に起因して抽出された文字列候補を除外し，残った候補をもとにモデル構築を行っている（山口 2017, 山口 2019）。残った候補のうち，変化率が減少傾向の例として「おるのであ」（図 8），増加傾向の例として「させていた」（図 9）を示す。

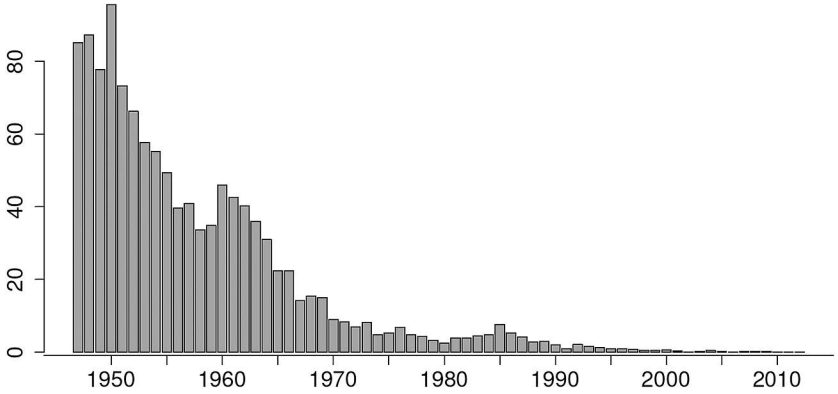

図 8 「おるのであ」

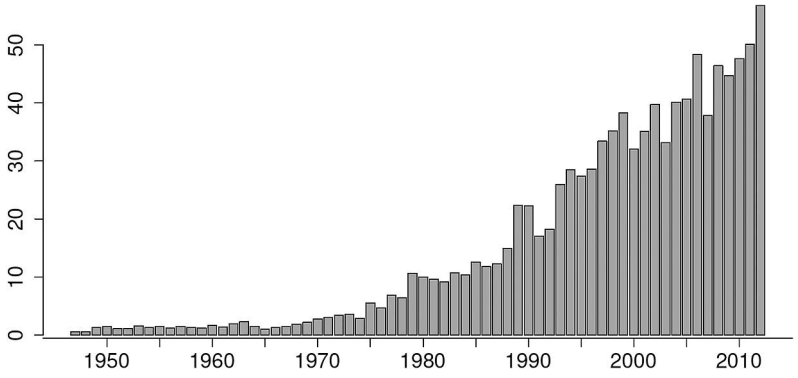

図 9 「させていた」

6. おわりに

本稿は，『国会会議録検索システム』に収録されている国会会議録のテキストデータに基づき，全文検索システム『ひまわり』用の『国会会議録』パッケージを構築する方法，および，構築結果を報告した。また，『国会会議録』パッケージを使って，(a) 経年変化が大きい表現を抽出する方法，(b) 抽出された表現に対する経年変化要因を調査する方法を示すことにより，『国会会議録』パッケージの有用性を示した。

参照文献

服部匡（2014）「現代日本語の通時変化」前川喜久雄（監修）・田野村忠温（編）『コーパスと日本語学』21–47. 東京：朝倉書店.

川瀬直人・清水茉有子（2015）「国会会議録フルテキスト・データベース Web API 開発の背景とその利用状況分析」『情報の科学と技術』65(12): 531–536.

国立国会図書館（2019）「国会会議録検索システム　検索用 API」https://kokkai.ndl.go.jp/api.html（2021 年 8 月 8 日確認）

松田謙次郎（編）（2008）『国会会議録を使った日本語研究』東京：ひつじ書房.

山口昌也（2017）「国会会議録における言語表現の時間的変化の予備的分析」『言語資源活用ワークショップ 2017』304–312.
山口昌也（2019）「国会会議録における言語表現の出現頻度に関する時間的変化モデルの検証」『言語資源活用ワークショップ 2019』321–329.
山口昌也・田中牧郎（2005）「構造化された言語資料に対する全文検索システムの設計と実現」『自然言語処理』12(4): 55–77.

Construction of a Language Resource Package of the Minutes of the National Diet of Japan for the Full-Text Search System "Himawari"

YAMAGUCHI Masaya

Spoken Language Division, Research Department, NINJAL

Abstract

This paper presents the method whereby a language resource package of the Minutes of the National Diet of Japan was constructed for the Full-Text Search System "Himawari" from text data stored in the Full-Text Database System for the Minutes of the Diet and reports the results of the construction. This package includes 11106 minutes (about 450 million characters) of the 1st (1947) to 182nd (2012) plenary sessions and budget committee meetings in the House of Representatives and the House of Councillors. Information related to the meetings, speakers, and the document structures of the minutes are annotated to the minutes in XML to facilitate the analysis of temporal changes in linguistic expressions. In this paper, I first describe the XML tags and an automatic annotation method created using notational clues in the minutes, then I detailed the application of the annotation method to the original minute data to construct the package and summarized the results. Finally, this paper classifies the usefulness of the package by showing how it can be used (a) to extract expressions showing large temporal changes and (b) to investigate the factors of the changes.

Keywords: The Minutes of the National Diet of Japan, language resource, full-text search system "Himawari", temporal change analysis