日本語コーパスとしての「国会会議録検索システム検索用API」
―計量的研究の精緻化・深化の可能性―

The Possible Applications of “Web API for Full-text Database System for the Minutes of the Diet” to a Corpus of Japanese

岡　田　祥　平
OKADA Shohei

１．はじめに

「国会会議録」を日本語研究資料と見なし，「国会会議録検索システム」（http://kokkai.ndl.go.jp/）を日本語研究に利用する可能性を提唱した松田謙次郎（2004）[1)]が発表されて，本稿執筆時点で既に10年以上が経過した。松田謙次郎（2004）以降，「国会会議録検索システム」を利用した日本語研究は，もはや珍しくなくなっている[2)]。ただ，「国会会議録検索システム」は日本語研究用に開発，公開されたものではないため，日本語研究に利用する際には留意しなければ点が存在することは，既に松田謙次郎（2004，2008，2012）などで指摘されている通りである。

本稿では，上述したような「国会会議録検索システム」を日本語研究として利用しようした種々の先行研究の試みを受けつつ，「国会会議録」を日本語研究，とりわけ計量的な研究を行おうとする際の新たな方法として，「国会会議録検索システム検索用API」（http://kokkai.ndl.go.jp/api.html。川瀬直人・清水茉有子2015）を使うことを提案することを目的とする。

本稿の構成は次のとおりである。まず，松田（2004，2008，2012）などに基づき，２節では日本語研究資料としての「国会会議録」の性格について，３節では「国会会議録」の検索システムである「国会会議録検索システム」を日本語研究に利用する可能性と限界について，それぞれ簡単に紹介する。４節では日本語研究用にデータが整備された「国会会議録」について概観したうえで，５節では「国会会議録」を利用した日本語研究の新たな手法として「国会会議録検索システム検索用API」を利用する選択肢を提案し，続く６節では「国会会議録検索システム検索用API」の利用方法について具体例をあげつつ概観する。最後に，７節で本稿のまとめとして，「国会会議録検索システム検索用API」をすれば，「国会会議録」を利用した日本語研究，特に計量的研究の精緻化・深化することが可能なのではないかと述べる。

なお，本稿では，これ以降，「国会会議録」「国会会議録検索システム」「国会会議録検索システム検索用API」という語を，それぞれ以下のような意味で使用することとする。

・「国会会議録」：
　国会での発言を文字化したデータ。「国会会議録検索システム」もしくは「国会会議録検索システム検索用API」の検索対象である。
・「国会会議録検索システム」（２節以降，「検索システム」と表記）：
　「国会会議録」を検索するためのシステムの一つ。http://kokkai.ndl.go.jp/ から利用可能。

2018.6.25　受理

・「国会会議録検索システム検索用API」（３節以降，「検索用API」と表記）：
「国会会議録」を検索するためのシステムの一つ。http://kokkai.ndl.go.jp/api.html から利用可能。

２．日本語研究資料としての「国会会議録」

前節で述べたとおり，「国会会議録」を日本語研究資料に利用する可能性を模索した嚆矢は，松田（2004）であると思われる。松田（2004）以降，精力的に「国会会議録」を日本語研究に利用する方法と可能性を模索し続けてきた松田氏は，日本語研究資料としての「国会会議録」の性格を，以下のように位置づけている（以下の７点は，松田2004，2008，2012の記述を，筆者がまとめたものである）。

① 1947年５月（第１回国会）から現在に至る約70年間における，
② 日本各地出身者の，
③ 19世紀後半生まれから20世紀後半生まれの成人（松田2004，2008は，「理論上は100年にわたる範囲の話者の発話を収めていることになる」としている）による，
④ 国会における発話という改まった場面での音声言語を，
⑤ 「できるだけそのままに記録しよう」とした，
⑥ 膨大な発話記録（山本和英2008によると，山本2008公表時点で約7.0GB・約35億文字，新聞記事コーパス約300年分）

以上の７点の詳細は松田（2004，2008，2012）にまとめられているのでそちらを参照していただきたいが，いずれにせよ，このような性格をもつ「国会会議録」は，日本語研究者，とりわけ（「現代」の範囲をどのように設定するかの厳密な議論はここではひとまず脇に置くことにして）「現代日本語」の動態に関心を持つ者にとっては，非常に魅力的な言語資料である（松田2004，2008には，「言語研究者であれば，当然国会会議録を日本語コーパスとして使うことを考えても不思議はない」という一節も存在する）。

もっとも，「国会会議録」は，日本語研究用に整備されたものではない。したがって，「国会会議録」を日本語研究に利用するには注意を要する点も，当然存在する。この点についても，松田（2004，2008）に基づいて簡単に紹介する。

まず，国会で発言されたものであっても，「国会会議録」には反映されないものが存在する，という点である。その代表的なものとしては，「議長の許可を得ない発言（「不規則発言」・やじ）」が挙げられる。「国会会議録」は議長の許可を得た発言を記録するものであるため，議長の許可を得ない「やじ」は「国会会議録」には反映されない。また，国会での「不穏当な発言」[3]は，議長の判断により「国会会議録」から取り消しされる。さらに，特殊な例としては，連合国軍占領下の1945年に出されたいわゆる「プレス・コード」（正式には「日本に与える新聞準則」）に抵触する発言も，「国会会議録」は反映されていない[4]。

次に，「秘密会」での審議は，「国会会議録」は存在していても公開はされていないという点も留意する必要がある。日本国憲法第57条第１項には「両議院の会議は，公開とする。但し，出席議員の３分の２以上の多数で議決したときは，秘密会を開くことができる」とあり，出席議員の議決の結果，「秘密会」となった審議についての「国会会議録」は参照できないわけである。

さらに，上にまとめた日本語研究資料としての「国会会議録」の性格の⑤に，「できるだけそのままに記録しよう」としたとあるが，「できるだけ」という留保がついている点にも留意が必要である。国会での発言は音声言語であるが，「国会会議録」は文字言語である。音声言語をそのまま文字化しても，文字言語としては意味の把握が困難であることは，少なくとも言語研究者であれば周知のことと思われるが，それゆえ，「国会会議録」も，国会での発言をそのまま文字化した状態で記録されているわけではない。速記録を日本語の表記に置きかえ，さらに整文（字句の整理）が施されているのである。ただ，「国会会議録」における整文の基準の詳細は公開されていないこともあり，国会での発言と「国会会議録」との間にどのような「ズレ」が生じているかは詳らかではない[5]。このような側面からも，「国会会議録」を日本語研究資料として利用するには，やはり注意を要するわけである[6]。

なお，「国会会議録」に格納されている日本語の性格については，茂木俊伸（2008）や松田（2012）の指摘に耳を傾ける必要があろう。

まず，茂木（2008）は，「国会会議録は，演説や質疑応答といった発言がいわば「地の文」となっているが，完全に同質のテキストによって構成されているわけではない」と指摘している。これは，「国会会議録」には「文書や法案の読み上げのような書き言葉の引用に近い発話」（茂木2008）なども含まれていることに対する警鐘である。したがって，「国会会議録」は「国会における発話という改まった場面での音声言語」（上記④）を格納していると言っても，その性格は一様ではないということには注意を払う必要があろう。

また，松田（2012）は，「国会会議録」には日本各地の出身者の発話が記録されているため，「国会会議録」を利用していわゆる「気づかない方言（気づかれにくい方言）」（当該話者は標準語／共通語だと認識しているが，実際は地域方言である語・表現のこと。詳細は，井上史雄1983，沖裕子1992などを参照）の研究も可能であると述べる一方で，以下のような警鐘を鳴らしている。

> ただし，バラエティがあると言っても，それは地域に限定した話です。社会的に見ると，国会議員は非常に偏った集団です。男女比，平均年齢，学歴，年齢などから考えても分かる通り，とても日本の縮図とはいえないどころか，かなり特殊な集団であると言えるでしょう。

勿論，国会では国会議員のみが発言しているわけではない。国会議員以外にも官僚や招致された参考人といった人々も発言をしている。しかし，国会における発言の大多数は国会議員によるものであり，「国会会議録」を日本語研究に利用する際には，以上の松田（2012）の指摘にも留意する必要があろう。

3．日本語研究資料としての「国会会議録検索システム」

日本語研究資料としての「国会会議録」の性格は前節で述べたとおりであるが，2001年より「国会会議録検索システム」（以下，「検索システム」と表記）の運用が開始されたことにより，オンライン上で「国会会議録」を検索することが可能になった。このことにより，「国会会議録」が日本語研究の資料として本格的に活用されるようになったわけであるが，「検索システム」を日本語研究に利用する際には以下に示すAからEのような点に注意をする必要があると，松田（2004，2008，2012）や茂木（2008）は指摘している。

A: OCRによる誤字・脱字
B: 外字処理
C: 表記のゆれ
D: 「検索件数」が検索対象語の用例数（出現度数）ではない
E: 検索の結果には，検索対象語の一部分しか一致しないものもヒットしてしまう

上記5点のうち，A「OCRによる誤字･脱字」とB「外字処理」は主に，紙媒体（官報号外）の「国会会議録」と電子化された「国会会議録」の相違点に当たり，電子化された「国会会議録」に存在する問題点である。（そして，本節の話題の中心である「検索システム」は，後者の電子化された「国会会議録」を検索することになる）。また，C「表記のゆれ」は紙媒体の「国会会議録」と電子化された「国会会議録」の双方に存在するが，「検索システム」を利用する際には無視できない問題である。さらに，D「「検索件数」が検索対象語の用例数（出現度数）ではない」やE「検索の結果には，検索対象語の一部分しか一致しないものもヒットしてしまう」は，「検索システム」利用上の問題点である。

以下，それぞれについて，簡単な説明を加える。

A「OCRによる誤字脱字」とは，紙媒体の「国会会議録」をOCRで電子化する際に生じる問題である。現在の「国会会議録」は紙媒体（官報号外）と電子データの双方の形式で作成されているが，実は，第145回国会（1991年1月開会）以前の「国会会議録」は紙媒体でしか作成されていない。したがって，「検索システム」で第145回国会以前の「国会会議録」を検索する場合は，紙媒体の「国会会議録」をOCRで電子化したデータを検索することになる。当然，紙媒体の資料をOCRで電子化する際には，OCRの誤認識による誤字・脱字が存在してしまうことになるが，前節の⑥で述べたとおり，「国会会議録」は膨大なデータ量のため，OCRの誤認識率がごく僅かであったとしても，少なくない誤字･脱字が発生してしまう。つまり，「検索システム」で特に第145回国会以前の「国会会議録」を検索する場合は，検索対象となるデータに少なくはない誤字・脱字が存在している，ということである。

B「外字処理」も，データの電子化と関係する問題である。すなわち，官報号外として印刷される紙媒体の「国

会会議録」では外字も使用されるが，「検索システム」の検索対象となる電子化された「国会会議録」では外字が使えないため，JIS 第 1・2 水準の字体に置き換える作業がされているという（松田2004，2008）。つまり，紙媒体（官報号外）の「国会会議録」と電子化された「国会会議録」には字体の異同が生じている場合がある，ということである。「検索システム」で検索を行う際には，この点についても留意しなければならないであろう。

C「表記のゆれ」は，同一語について複数の表記が観察される問題である。この問題は，「検索システム」を外来語研究に利用する可能性を模索した茂木（2008）で指摘されているが，「検索システム」を利用にするに当たって「表記のゆれ」が大きな問題になるのは，以下の二つの理由による。

C-1　どのような表記のゆれのパターンが存在しているかがわからない

C-2　「検索結果に影響が出る（検索件数が異なる）場合と出ない場合（検索件数がおなじになる場合）がある」（茂木2008）

「国会会議録」にどのような表記のゆれのパターンがあるかというC-1の問題については，検索対象語について，考え得るあらゆる表記のパターンを検索することでなんとか対応できると考えられる。一方，C-2の問題は，問題を提起した茂木（2008）が述べているように，「検索システムの仕様が一般公開されていない現状では，検証を行いながら経験的に理解せざるを得ない」しか，対応策はなさそうである。

D「「検索件数」が検索対象語の用例数（出現度数）ではない」という問題は，「検索システム」で検索した結果，表示される「検索件数」は国会における検索対象語の用例数（出現度数）ではない，ということである。実は，「検索システム」が表示する「検索件数」は，検索対象語を含む会議の数なのである。したがって，「1 つの会議で，複数の発言者が，あるいは 1 人の発言者が複数回，ある語を使用していたとしても，「検索件数」としては「1」になって」（茂木2008）しまうのである。このことは，ある語がある会議で1,000回使用されていたとしても，「検索システム」の「検索件数」は 1 となってしまうということを意味しており，場合によっては計量的な研究を行う場合に深刻な影響を与えてしまう[7)]。

E「検索の結果には，検索対象語の一部分しか一致しないものもヒットしてしまう」というのは，たとえば，検索したい語は「スクリーニング」であるにもかかわらず，検索結果の中に「ハウスクリーニング」という語も含まれてしてしまう（茂木2008）という問題である。さらに厄介なのは，「検索システム」は検索の際に句読点を無視するために，松田（2012）が指摘するように，「言語学」を検索した場合に，「生活言語，学校言語」と，読点を超えて，この種の問題が生じてしまうという点である。このように，「検索システム」の「検索件数」には，場合によっては相当数，検索対象語以外の語が含まれてしまうのである。

以上，述べてきたA，B，C-1の問題は，基本的には「国会会議録」（の電子化データ）を作成する時点での問題であり，「検索システム」のユーザーには対応が困難な問題である。しかし，D「「検索件数」が検索対象語の用例数（出現度数）ではない」という問題と，E「検索の結果には，検索対象語の一部分しか一致しないものもヒットしてしまう」という問題は（そして，C-2「検索結果に影響が出る」表記のゆれがあるという問題も），「国会会議録」自体の問題ではない（「検索システム」の問題である）ためユーザー側でも対応が可能ではあるし，何より，「国会会議録」を日本語研究，とりわけ計量的な研究に利用する場合にはなんとかして対応しなければならない問題である。

ただ，DとE（とC-2）の問題が「検索システム」のユーザーで対応可能ではあるが，問題解決の方策はなかなか厄介である。

この二つ（あるいは三つ）の問題を解決するために，たとえば茂木（2008）は，「検索結果をすべてテキストファイル化し，人手で用例数に直す」という作業を行っている。ただ，茂木（2008）がどのようにして「検索結果をすべてテキストファイル化」したのか不明である（「検索システム」は検索結果をダウンロードできるが，それは検索の結果，ヒットした会議を一つ一つ手作業でダウンロードしなければならず，「検索件数」が膨大である場合，そのような作業を行うのは莫大な時間と手間を要する）。また，「検索件数」が膨大である場合，「人手で用例数に直す」にどの程度の労力が要するのかも想像できない。[8)]

４．日本語研究用にデータが整備された「国会会議録」

前節では,「国会会議録」ならびに「検索システム」を日本語研究に利用する際の可能性と問題点について,種々の先行研究での指摘を概観した。そのうち，問題点については，結局は「国会会議録」ならびに「検索システム」が日本語研究用に整備されていない故に生じるものであるといえる。そのように考えると,「国会会議録」を利用した日本語研究を行うには，日本語研究用にデータが整備された「国会会議録」が利用できるのが理想的である。

実は，日本語研究用に整備された形の「国会会議録」として，現在，以下の２種類が利用可能である。

・全文検索システム『ひまわり』用の「国会会議録」パッケージ

・『日本語書き言葉均衡コーパス』の「国会会議録」データ

『ひまわり』とは「言語研究用の全文検索システム」（山口昌也2013）のことで，国立国語研究所のページ（http://www2.ninjal.ac.jp/lrc/）から入手できる。上述のリンク先からは，『ひまわり』用に情報が整備された「国会会議録」のデータ（形態素解析済み・発話者の生年代の情報も付与）も入手可能であり，これを利用すれば，効率よく「国会会議録」を利用した日本語研究が行える。ただ,『ひまわり』用の「国会会議録」パッケージは, 1947年から2012年までの, 衆参両議院の本会議と予算委員会のデータのみの提供であり，２．節で述べた日本語研究資料としての「国会会議録」の性格の特徴のうち, 主に①「1947年５月（第１回国会）から現在に至る約70年間における」という点と，⑥「膨大な発話記録」という２点について，ある種の制約が生じてしまう。

一方,『日本語書き言葉均衡コーパス』（BCCWJ・The Balanced Corpus of Contemporary Written Japanese）とは，国立国語研究所によって構築された「現代日本語の書き言葉の全体像を把握するために構築したコーパスであり，現在，日本語について入手可能な唯一の均衡コーパス」（http://pj.ninjal.ac.jp/corpus_center/bccwj/）のことである。BCCWJの利用方法などについては上述のリンク先などを参照いただきたいが，BCCWJでは，第77回国会（1976年）から第163回国会（2005年）までの30年間分の「国会会議録」が, 形態論情報（http://pj.ninjal.ac.jp/corpus_center/bccwj/morphology.html などを参照）などが付与された状態で利用可能である。ただ，上述の通り，BCCWJに格納されている「国会会議録」は1976年から2005年までの30年間と限定されている上に,「両院協議会で開かれた61会議，発言部分の文字数が1000文字以下の6401会議，第77回国会のうち1975年に開催された33会議は除外」（丸山岳彦・柏野和佳子2014）されている。つまり,『ひまわり』用の「国会会議録」パッケージ同様，BCCWJに格納されている「国会会議録」のデータも，２．節で述べた日本語研究資料としての「国会会議録」の性格の特徴のうち，主に①「1947年５月（第１回国会）から現在に至る約70年間における」という点と，⑥「膨大な発話記録」という２点について，ある種の制約が生じてしまっているのである。

松田（2014）は,「本来は国会の記録のために作られている」「「国会会議録」を，言語学者が言わば勝手に研究に使いだしたわけですから，少々の不自由があるのはやむを得ないと言うべき」と述べている。確かに松田（2014）の指摘する通りなのであるが，２．節で述べた日本語研究資料としての「国会会議録」の性格の特徴すべてを活かす形―すなわち,「国会会議録」全体を対象に検索する形―で，なおかつ（比較的）簡便な方法で「国会会議録」を日本語コーパスとして利用できる方法がないか, 考えていた筆者の目に止まったのが，次節で紹介する「国会会議録検索システム検索用API」である。

５．「国会会議録検索システム検索用API」の概要－日本語研究への応用可能性を意識しつつ―

「国会会議録検索システム検索用API」とは,「国会会議録検索システムに登録されているデータを検索し，取得するための外部提供インターフェイス（API：Application Programming Interface）」（http://kokkai.ndl.go.jp/api.html より。以下,「検索用API」と表記）のことで，次節で詳述するが，指定されたURLに検索条件を付与しHTTPの「GETリクエスト」として送信すると，検索条件に合致する「国会会議録」をXML文書（6.3節の図２, 6.4節の図５・図６も参照）として戻してくれるもので, 2014年12月に公開されている（詳細は，川瀬・清水2015も参照）。

「検索用API」で指定できる検索条件は,「検索システム」の「簡単検索」で指定可能な条件と同様, ①期間, ②発言者名, ③院名, ④会議名, ⑤検索語の五つである。つまり,「検索用API」は, 検索条件, 機能の側面からは,「検索システム」の「簡単検索」同等の機能を持つということになる。

なお,「検索用API」での検索の結果, 取得できるXML文書は, 以下の二つの形式から選択できる。

（Ⅰ） 発言単位： 検索条件に合致する「発言」を抽出

（Ⅱ） 会議単位： 検索条件に合致する「会議録」（検索条件に合致する「発言」を含む, 会議全体の記録）を抽出

検索結果を「発言単位」で得るか, それとも「会議単位」で得るかは, 目的による。この点については, 川瀬・清水（2015）の以下の記述も参考にされたい。

> 例えば, ある特定の語を含む発言, あるいはある特定の発言者だけを取り上げたいというような場合には前者（引用者註：「発言単位」のこと）を利用するのが適当である。そうではなく, 検索した語についてどのような議論が行われたか, という会議録としての議論の文脈を見る必要がある場合には, 会議単位を使うことで全体の議論を取得することができる。

さて,「検索用API」が「検索システム」の「簡単検索」同様の機能を持つならば, 国会会議録を利用した日本語研究を行うにあたって, わざわざ「検索用API」を利用しなくとも, 以前より利用されてきた「検索システム」を利用し続ければいいのではないか, という声があるかも知れない。しかし, 筆者は, 上述の通り,「検索用API」を利用すれば, 検索結果を半ば自動的にXML文書で入手できる（詳細は次節で説明する）という理由から,「検索用API」を日本語研究資料として積極的に活用しても良いのではないかと考えている。

もっとも,「国会会議録」の検索結果をダウンロードできるというのは, 何も「検索用API」だけで利用可能な機能というわけではない。３節の末尾でも少し触れたとおり, 従来の日本語研究で積極的に利用されてきた「検索システム」でも検索結果をダウンロードできる（松田2008も参照）。ただ, これも３節の末尾でも触れたとおり,「検索システム」で検索結果をダウンロードしようとする場合, その作業は検索結果１件１件について個別に手作業で行わなければならず, 特に検索結果が多い場合, すべての検索結果をダウンロードするのは非常に手間がかかるのである（コンピュータ技術に長けている場合は何らかの方法で検索結果を自動的にダウンロードすることもできるのかもしれないが, 少なくとも筆者は大変恥ずかしいことにその術を知らない）。一方,「検索用API」は, 既に繰り返し述べているように, 検索条件に合致する結果を半ば自動的にXML文書の形で得ることができる。このことは, 検索結果を様々な方面（検索対象語におけるコロケーション研究など）に活用することが容易になると同時に, ３節で指摘した「検索システム」がもつ二つの問題点―すなわち, D「「検索件数」が検索対象語の用例数（出現度数）ではない」という問題と, E「検索の結果には, 検索対象語の一部分しか一致しないものもヒットしてしまう」という問題―は, ７節で実例を紹介しながら説明するように,「検索用API」を利用すれば（簡単に）解決できるのである。つまり,「国会会議録」を日本語研究資料とした多くの研究が利用してきた「検索システム」では難しかった正確な用例数をもとに議論, 研究が,「検索用API」を利用すれば, 比較的簡単に行える, ということである。

ただし,「検索用API」にも短所は存在する。その大きなものは,「検索用API」を利用すれば検索条件に合致する結果を半ば自動的にXML文書の形で得ることができるものの,「検索用API」の１回の検索で取得可能なXML文書は,「発言単位」で100件,「会議単位」で５件しか得られない, という点である。つまり, 検索結果の件数が多い場合はXML文書を取得する操作を繰り返す必要がある。詳しくは６．４節で説明するが, たとえば検索条件に合致する検索結果が１万件であった場合, XML文書を取得する操作を「発言単位」検索では100回,「会議単位」の検索では実に2,000回, 繰り返さなければならないのである。この事実をどのように「解釈」するかは個人差があるであろうが, 上述したとおり,「検索システム」を利用した場合であれば検索結果を１件１件手作業でダウンロードしなければならず（つまり, 検索結果が１万件の場合, １万回の作業を要する）, その点を踏まえるならば,「検索用API」を利用することは, 特に研究者の負担が軽減することにつながると筆者は考えている（特に「発言単位」での検索の場合）[9]。

なお, 川瀬・清水（2015）は,「検索用API」の活用が見込まれる分野, 方向性について,「日本語言語コーパスとしての活用」「特定の主題や政策研究における活用」[10]「政策情報公開への活用」[11]「国会活動の量的な可視化」[12]の四つを上げている[13]。中でも,「日本語言語コーパスとしての活用」の可能性に関連し, 川瀬・

清水（2015）は以下のように述べている。

> 国会会議録は戦後すぐの時点から現在まで続く長期間のデータであり，相当量のデータを持ったデータベースであることから，日本語学研究や自然言語処理などの分野において利用価値のあるデータベースと考えられている[14]。本 API を利用することによって，テキストデータの取得と形態素解析等を用いた分析とを結びつけることが容易になるため，この種の研究での活用が進むことが期待される。

ただし，「国会会議録の API 利用はまだ低調である」（川瀬・清水2015）という。川瀬・清水（2015）が発表されて約２年半が経過した本稿執筆時点（2018年７月）においては，「検索用API」の利用状況が変化したかも知れないが，川瀬・清水（2015）が「検索用API」を「日本語言語コーパスとしての活用」可能性を提言しているにもかかわらず，少なくとも筆者の目に止まった範囲では，「検索用API」を「日本語言語コーパスとしての活用」した研究は，筆者の試み（註13も参照のこと）以外に存在していないようである。それゆえ，「検索用API」が日本語研究資料として活用されることを願い，「検索用API」の開発者でもなく，また，「国会会議録を使った日本語研究」（松田［編］2008）を利用した積極的に展開したわけでもない筆者であるが，あえて本稿を執筆しようと思うに至った次第である。

６．「国会会議録検索システム検索用API」の利用法

前節では，「国会会議録」を日本語研究資料として活用する際には「検索用API」を利用する選択肢もあり得るが，「検索用API」を活用した日本語研究はほとんど存在していないという２点を述べてきた。「検索用API」を利用した日本語研究がほとんど存在していない理由としては，日本語研究者に「検索用API」の存在がまだ知られていない可能性と同時に，日本語研究者に「検索用API」の存在が知られていたとしてもその利用法が知られてない可能性も考えられる。というのも，「検索用API」の利用方法は，「検索用API」のホームページや川瀬･清水（2015）などを確認しても，ごく簡単にしか説明されておらず，特にコンピュータリテラシーが高くない筆者のような人間にとっては，「検索用API」を利用するには，若干，心理的な抵抗を感じてしまう側面があるのも，否定できないからである。そこで，本節では，現代日本語研究に関心を持つ方に「検索用API」を日本語研究資料，特に日本語コーパスとして使っていただくことを念頭に置いて，自身の非力を承知しつつ「検索用API」の利用方法を筆者なりに説明していく（「筆者なりに」と書いたのは，「検索用API」のホームページや川瀬・清水2015には説明されておらず，筆者が「検索用API」を自ら利用するなかで「発見」した事柄も含まれるからである）。

６．１　「検索用API」の利用にあたっての基本事項

「検索用API」を利用するにあたっては，インターネットに接続しているパソコン，およびインターネットブラウザ（ウェブブラウザ）が必要である。このように書くと小難しく思えるが，要はインターネットが閲覧できる環境にあれば問題はない。そして，「検索用API」を検索するにあたっては，以下に示すような検索条件を入力したURLをインターネットブラウザに入力して，接続するだけで良い（「発言単位」「会議単位」については，５節での説明を参照）。

・「発言単位」で検索したい場合：　http://kokkai.ndl.go.jp/api/1.0/speech?｛検索条件を入力｝
・「会議単位」で検索したい場合：　http://kokkai.ndl.go.jp/api/1.0/meeting?｛検索条件を入力｝

上記URLをご覧いただくとおわかりいただける通り，「検索用API」を活用するためには，「｛検索条件を入力｝」という部分に，自分が検索したい条件を指定された形式に従って入力できることが肝要になる。次節ではその方法を説明する。

６．２　「検索用API」で指定可能な検索条件

「検索用API」で指定可能な検索条件は，以下の表１に示すとおりである。

表１で「必須区分」が「いずれか必須」となっている「院名」「会議名」「検索語」「発言者名」「開会日付／始点」「開会日付／終点」のすべてを省略して検索用のURLを作成した場合は，エラーとなり，検索結果は戻されない。

また，「会議名」「検索語」「発言名」においては，各検索条件を半角スペースでつなぐことで，複数条件を指定した検索が可能である。ただし，検索方法（AND ／ OR検索）は固定で変更できない。つまり，「会

議名」「検索語」「発言名」においては複数条件を指定した検索が可能であるものの，「会議名」「発言者名」についてはOR検索しかできず，「検索語」についてはAND検索しかできない，ということである（AND／OR検索については次節での議論も参照）。

表１のうち，「院名」「会議名」「検索語」「発言者名」「開会日付／始点」「開会日付／終点」は（「検索用API」だけではなく）「検索システム」の「簡単検索」でも指定可能な条件である。一方,「開始位置」と「一回の最大取得件数」は「検索用API」独自の検索条件である（検索結果をXML文書で入手する際に関係する条件である）。

なお，表１で示されている「パラメータ名」は，検索URLを作成する際には大文字／小文字の区別をする必要はないようである。

表１　「検索用API」で指定可能な検索条件一覧

（http://kokkai.ndl.go.jp/api.html をもとに筆者が作成）

項目名	必須区分	パラメータ名	複数指定	入力制限（※）	備考
開始位置	任意	startRecord	不可	10バイト以内	総結果件数の中の出現位置を指定する。 省略時のデフォルト値は「発言単位」，「会議単位」ともに「1」
一回の 最大取得件数		maximumRecords	不可	10バイト以内	「発言単位」で検索する場合は，1回のリクエストで取得できる発言の数 「会議単位」で検索する場合は，1回のリクエストで取得できる会議の数 省略時のデフォルト値は「発言単位」では「30」，「会議単位」では「2」 上限値は発言単位では「100」，会議単位では「5」
院名	いずれか必須	nameOfHouse	不可	20バイト以内	完全一致検索（衆議院，参議院，両院，両院協議会のいずれかを指定可能） 指定しない場合は全院で検索 「両院」と「両院協議会」はどちらも指定可能で，検索結果は同一
会議名		nameOfMeeting	可（OR検索）	1000バイト以内	部分一致検索
検索語		any	可（AND検索）	256バイト以内	部分一致検索
発言者名		speaker	可（OR検索）	1000バイト以内	部分一致検索
開会日付／始点		from	不可	10バイト以内	YYYY-MM-DD で指定 省略時のデフォルト値は「0000-01-01」
開会日付／終点		until	不可	10バイト以内	YYYY-MM-DD で指定 省略時のデフォルト値は「9999-12-31」

※ UTF-8で送信されたリクエストは内部的にEUCに変換される。
各パラメータの入力制限は文字コードをUTF-8からEUCへ変換した後のバイト数で，複数条件指定時に用いるスペースもバイト数に含む。
検索条件全体では，UTF-8をURLエンコードした状態で2000バイトが上限。

６．３　「検索用API」の検索URL

では，「検索用API」で条件を指定して検索する場合，具体的にどのようなURLを作成すればよいのだろうか。以下，いくつか実例を紹介しながら，確認していきたい（適宜，表１も参照されたい）。

(1)　http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=1&maximumRecords=5&any=アベノミクス&speaker=安倍晋三

(1)のURLが意味するところは，以下の通りである（二重下線の意味は後述。なお，(1)は検索用APIのホームページに掲載されている例である）。

(1)'　speech　：　「発言単位」で検索する
　　startRecord = 1　：　検索結果の通し番号の１番目から検索結果をXML文書で戻す
　　maximumRecords = 5　：　５件分の検索結果を戻す
　　any = アベノミクス　：　検索対象語は「アベミクス」
　　speaker = 安倍晋三　：　発言者は「安倍晋三」

(1)'を見ると分かるように，それぞれの検索パラメータ名と検索者が指定したい検索条件を「（半角の）=」を利用して結びつける。また，(1)を見ると分かるように，それぞれの検索パラメータは「（半角の）&」で結びつける。

ただし，(1)のURLのインターネットブラウザに入力して検索を実行しても，「(19011)検索条件の入力に誤りがあります。」というエラーメッセージが戻ってくる。それは，「検索用API」の検索URLを作成する際には検索者が入力した検索条件（(1)の二重下線部）をURLエンコードする必要があるのだが，(1)で示しているURLの状態ではエンコードがされていないため，検索できないのである。

「検索用API」の検索URLを作成する際には，以下の(2)で示した諸要素を，UTF-8でURLエンコードをする必要がある[15]。(1)であれば，二重下線を付した部分がURLエンコードをする必要がある部分である。

(2)　検索パラメータ入力時に使うもの：　「（半角の）=」「（半角の）&」「半角スペース」
　　検索条件時に入力するもの：　ひらがな　カタカナ　漢字

さて，「UTF-8でURLエンコードをする必要」と聞くと，戸惑われる方もいらっしゃるかも知れないが，実際は非常に簡単である。「URLエンコード」を検索語にして検索エンジンで検索を行えば，適当なサイトが複数ヒットする。その中で，任意のサイトを利用してURLエンコードをすればいいだけである。なお，筆者が使用を推奨するのは，以下のサイトである（推奨する理由は後述）。

(3)「URLのエンコード・デコード - 日本語文字コード」： http://charset.7jp.net/urlchg.html

(3)で示したようなサイトを利用し，対応する必要がある部分のURLエンコードを行う。たとえば，「any=アベノミクス」を(3)に紹介したサイトを利用してUTF- ８でURLエンコードを行うと，以下の図１のようになる（「検索用API」ではUTF- ８でURLエンコードを行う必要があるので，「URLのエンコード・デコード - 日本語文字コード」＝図１の「文字コード」は「UTF- ８」を選択した）。

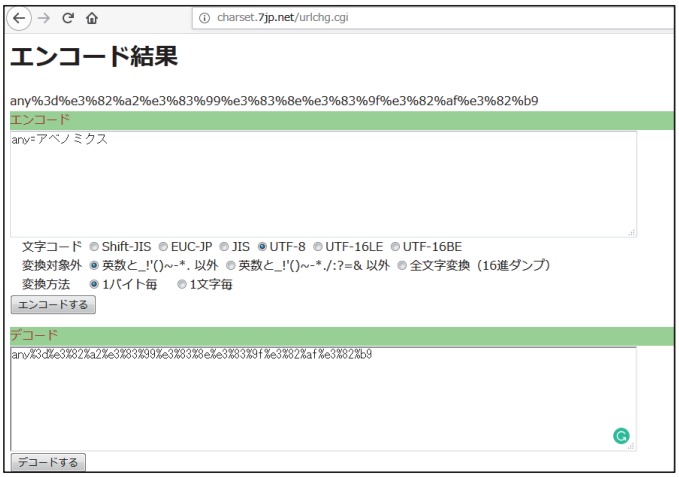


図１　URL エンコードの一例①

図１でいうと，「デコード」欄に表示されたものを検索の際のURLに利用すれば良い。

さて，(1)の例について，必要な箇所のURLエンコード行うと，以下の(1)”のようなURLになる。

(1)” http://kokkai.ndl.go.jp/api/1.0/speech?startRecord%3d1%26maximumRecords%3d5%26any%3d%e3%82%a2%e3%83%99%e3%83%8e%e3%83%9f%e3%82%af%e3%82%b9%26speaker%3d%e5%ae%89%e5%80%8d%e6%99%8b%e4%b8%89

(1)”のURLをインターネットブラウザに貼り付け，エンターキーを押し，検索をすれば，以下の図２のような検索結果が得られる。

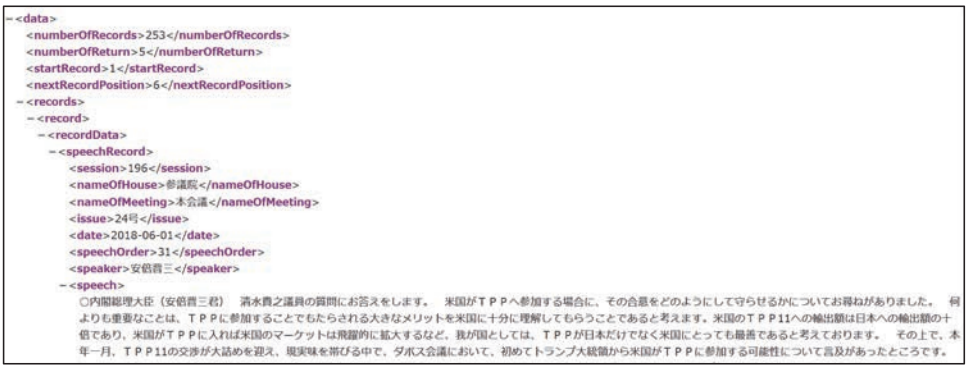


図２　「検索用 API」の検索結果

図２のような検索結果が得られれば成功であるが，はじめはうまくいかない場合も少なくない。その原因としてはURLエンコード以前の段階（(1)のような状態）で検索条件を正しく入力していないという場合も多いが，意外と多いのは，URLエンコードをする部分を間違えているというケースである。具体的には，(1)でいうと，二重下線が付されている部分を過不足なくURLエンコードをする必要があるのだが，(1)を見ても分かる通り，URLエンコードをすべき部分とURLエンコードをしなくてもよい部分が混在しており，URLエンコードをする時点で失敗が起きやすい（URLエンコードをしなければいけない部分をしなかったり，逆にURLエンコードをしなくてもいい部分をしてしまったりするというミスを犯しがちである，ということである）。

しかし，URLエンコードを行う際，(3)で紹介した「URLのエンコード・デコード - 日本語文字コード」を利用すれば，検索条件として入力した箇所を一括して適切な形でURLエンコードを行うことができる。以下，引き続き，(1)を例に上げ，その方法を説明する。

(1)の場合，検索者が検索条件として指定しているのは，「startRecord」以降の部分である。そこで，「startRecord」以降の部分をコピーし，「URLのエンコード・デコード - 日本語文字コード」の「エンコード」欄（図１でいうと上の欄）に貼り付ける。そのうえで，「URLのエンコード・デコード - 日本語文字コード」の画面で，以下のような設定を行い（図２も参照），「エンコードする」ボタンを押す。

(4)「文字コード」　：　「UTF- ８」を選択
「変換対象外」　：　「英数と_!'()~-*. 以外」を選択
「変換方法」　：　「1バイト毎」を選択

そうすると，図３のような結果が得られる。

図３　URL エンコードの一例②

以上のようにして得られた結果（図３の「デーコド」欄の結果）を,「発言単位」の検索を行うURL（http://kokkai.ndl.go.jp/api/1.0/speech?）のうしろに貼り付け検索をかけると，図２のような結果が得られるというわけである。

筆者の少ない経験で恐縮なのだが,「検索用API」でURLエンコードをする際，検索条件ごとに個別に作業を行うとミスを犯しがちであったので，読者のみなさまには「URLのエンコード・デコード - 日本語文字コード」を利用し一括してURLエンコードを行うことをお薦めする次第である。

さて，(1)で示した以外の条件を指定した検索を行う場合は，どのようにすればよいであろうか。以下，いくつかの例を上げておく。

(5)　http://kokkai.ndl.go.jp/api/1.0/speech?nameOfHouse＝衆議院＆nameOfMeeting＝予算委員会 本会議＆any＝多言語 日本語

(5)の検索URLの意味は，以下の通りである（勿論，(5)の場合も，検索を行う際には，「startRecord」以降の二重下線が付されている部分をURLエンコードする必要がある）。

(5)' speech : 「発言単位」で検索する
nameOfHouse = 衆議院 : 検索対象の議会は「衆議院」
nameOfMeeting = 予算委員会 本会議 : 検索対象は「予算委員会」もしくは(or)「本会議」
any = 多言語 日本語 : 検索対象語は「多言語」と(and)「日本語」

(5)の検索URLで留意しなければいけないのは，複数の条件を指定している検索パラメータがあることと，検索パラメータによって指定した複数の条件の「意味合い」が異なるという点である。

(5)'を見ても分かる通り，(5)の検索URLでは，検索パラメータのうち「会議名」（nameOfMeeting）と「検索語」（any）で複数の検索条件を指定している。ここで思い出していただきたいのは，表1に関連して説明した，以下の事項である。

(6)① 「会議名」「検索語」「発言名」においては複数条件を指定した検索が可能である
② ただし，「会議名」「発言者名」についてはOR検索しかできない
③ 「検索語」についてはAND検索しかできない
④ ②・③の設定は，検索者の意向によって変更できない

「検索用API」は(6)のような仕様であるため，(5)の検索URLの場合，「会議名」は検索条件として指定した「予算委員会」と「本会議」で【OR検索】を実行する（「予算委員会」と「本会議」，どちらか一方でも検索条件にヒットするものを検索結果として戻す）一方で，「検索語」では検索条件として指定した「多言語」と「日本語」で【AND検索】を行うことになる（「多言語」と「日本語」の双方を含む発言を検索結果として戻す）。

「検索用API」で複数条件が指定できる検索パラメータについて，AND／OR検索の仕様を検索者の意向によって変更できない点は，日本語研究に「検索用API」を利用する際には，若干不便な面も否めない。というのも，日本語研究の場合（少なくとも筆者は），特に「検索語」については調査目的に応じて「OR検索」と「AND検索」を使い分けたくなるからである。このような事情を踏まえると，「検索用API」で複数の条件を指定して検索を行う際，場合によっては，その都度，検索URLを別個に作成して，それぞれの条件において検索するほうが，結果的には負担が少ないのかも知れない。

さて，次に(7)の検索URLの例を見ていくことにしよう。

(7) http://kokkai.ndl.go.jp/api/1.0/meeting?startRecord=1&maximumRecords=5&any=日本手話 日本語対応手話&from=2001-01-01&until=2017-12-31

(7)の検索URLの意味は以下の通りである（言うまでもないことであるが，(7)の場合も，検索を行う際には，「startRecord」以降の二重下線が付されている部分をURLエンコードする必要がある）。

(7)' meeting : 「会議単位」で検索する
startRecord=1 : 検索結果の通し番号の1番目から検索結果をXML文書で戻す
maximumRecords=5 : 5件分の検索結果を戻す
any= 日本手話 日本語対応手話 : 検索対象語は「日本手話」と(and)「日本語対応手話」
from=2001-01-01 : 検索開始日は「2001年1月1日」
until=2017-12-31 : 検索終了日は「2017年12月31日」

なお，検索対象期間を指定する際の日付は必ず4桁である必要がある（(7)の例であれば，「from=2001-1-1」というURLで検索しようとするとエラーとなってしまう，ということである）。

以上，「検索用API」の検索URLに関する説明を簡単に行ってきたが，筆者の説明能力にも問題があり，要領を得ない読者もいらっしゃることかも知れない。ただ，「検索用API」における検索URLの作成方法は，「検索用API」のホームページに書かれている説明や，川瀬・清水（2015），さらには筆者による上記の説明で可能なはずである。あとは，読者のご関心にしたがって，種々の検索を行っていただくことで，「検索用API」を利用した検索技術を高めていただければと思っている（なお，次節以降にも検索URLの具体例を提示しているので，そちらも参照いただきたい）。

6.4 「検索用API」での検索結果が多い場合の対応

「検索用API」では，前節まで述べてきたような検索URLに対して，図2で示したようなXML文書で検索結果が戻される。このXML文書をどのように利用するかについては6.5節や7節で述べるが，そもそも，

「検索件数」が多数の場合，「検索用API」での初期設定ではXML文書で全ての検索結果を戻してくれないということを承知しておく必要がある。このような場合は，検索URLを操作し，「検索件数」を押さえるという対応を取らなくてはならない。本節では，そのような場合の対応策について，筆者の試み（岡田祥平2018b）を例に取り上げつつ，紹介したい。

岡田（2018b）では，第二次世界大戦後の国会（1947年５月開会の第１回国会から，2017年12月31日まで）において，「国語」「日本語」という語がどのように使用されてきたのか，「計量テキスト分析」（詳細は樋口耕一2014を参照）の手法を利用し，考察を試みたものある。「計量テキスト分析」を行うにあたっては，「国会会議録」から「国語」「日本語」を含む発言を取り出す必要があるが，その際に筆者は「検索用API」を利用したというわけである。具体的な手順は以下のとおりである。

まず，「検索用API」を利用し，第二次世界大戦後の国会において，「国語」という語が含まれるすべての発言を取り出すために，(8)のような検索URLを作成し，実行した（二重下線部はURLエンコードが必要な部分。ただし，これ以降，煩雑さを避けるため，本文中においてはURLエンコードを施した結果は提示しない）。

(8) http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=1&maximumRecords=100&any=国語

(8)の検索URLの意味は，以下の通りである。

(8)' speech　：　「発言単位」で検索する
startRecord=1　：　検索結果の通し番号の１番目から検索結果をXML文書で戻す
maximumRecords=100　：　100件分の検索結果を戻す
any=国語　：　検索対象語は「国語」

(8)の検索URLを実行すると，「検索件数」がシステムの上限（1,000件）を超えているため，以下のようなエラーメッセージが戻ってくる（ただし，「検索件数」の上限が1,000件というのは名目上のことで，以下の図５や(10)で示す通り，実際は「検索件数」が1,000件を多少超える程度であれば，問題はないようである）。

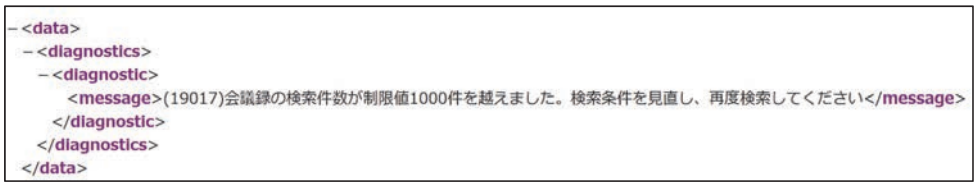


図４　「検索用 API」で「検索件数」が上限を超えたというエラーメッセージの例

検索結果をXML文書で入手するには，(8)で指定した検索条件にさらなる条件を加え，「検索件数」を減らす必要がある。そのための方法としてはいくつかの可能性が考えられるが，筆者は検索対象期間を限定するという対応をとることにし，以下の(9)のような検索URLを作成した。なお，斜体字にした部分が(8)の検索URLに，さらに新たに加えた条件である。

(9) http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=1&maximumRecords=100&any=国語&*from=1947-05-01*&*until=1960-12-31*

(9)の検索URLの意味は，以下の通りである。

(9)' speech　：　「発言単位」で検索する
startRecord=1　：　検索結果の通し番号の１番目から検索結果をXML文書で戻す
maximumRecords=100　：　100件分の検索結果を戻す
any=国語　：　検索対象語は「国語」
from=1947-05-01　：　検索開始日は「1947年５月１日」
until=1960-12-31　：　検索終了日は「1960年12月31日」

すると，今度は図５のような結果が戻されてくる。つまり，(9)の検索URLでの検索は成功したということである。

図５から読み取れる種々の情報は，以下の通りである。

(10) ２行目「<numberOfRecords>1445</numberOfRecords>」:
(9)の検索URLの条件を満たす「検索件数」は1,445件である
３行目「<numberOfReturn>100</numberOfReturn>」:
XML文書として戻している発言の件数が100件である
４行目「<startRecord>1</startRecord>」:
検索結果のうち，１番目のものからXML文書を戻している
５行目「<nextRecordPosition>101</nextRecordPosition>」:
次の検索結果は，101番目から始まる

```
<data>
<numberOfRecords>1445</numberOfRecords>
<numberOfReturn>100</numberOfReturn>
<startRecord>1</startRecord>
<nextRecordPosition>101</nextRecordPosition>
<records>
<record>
<recordData>
<speechRecord>
<session>37</session>
<nameOfHouse>衆議院</nameOfHouse>
<nameOfMeeting>内閣委員会</nameOfMeeting>
<issue>5号</issue>
<date>1960-12-22</date>
<speechOrder>37</speechOrder>
<speaker>西村直己</speaker>
<speech>
○西村国務大臣　自衛隊の基本の問題でありますから、私から申し上げます。御存じの通り自衛隊の基本のあり方は、自衛隊法の
```


図５　「検索用 API」での検索結果①

(10)の内容をまとめるならば，「(9)の検索URLで得られた検索結果は1,445件であるが，今回示したXML文書（図５）はそのうち，１番目の検索結果から100番目の検索結果をXML文書として表示している」ということになろう。したがって，(9)の検索URLの条件を満たす1,445件のXML文書全てを入手するには，引き続き，作業を行う必要がある。具体的には，以下の(11)のように，(9)の検索URLの「startRecord=1」を「startRecord=101」にして検索を行うだけでよい（斜体字が(9)から変更した部分）。すなわち，図５の５行目に「<nextRecordPosition>101</nextRecordPosition>」（＝次の検索結果は，101番目から始まる）と表示されているので，検索URLを「startRecord=101」（＝検索結果の通し番号の101番目から検索結果をXML文書で戻す）という指示をすればいい，ということである。

(11) http://kokkai.ndl.go.jp/api/1.0/speech?startRecord =*101*&maximumRecords =100&any =国語 &from =1947-05-01&until =1960-12-31

(11)の検索URLを実行すると，以下の図６のようなXML文書が戻される。

```
<data>
<numberOfRecords>1445</numberOfRecords>
<numberOfReturn>100</numberOfReturn>
<startRecord>101</startRecord>
<nextRecordPosition>201</nextRecordPosition>
<records>
<record>
<recordData>
<speechRecord>
<session>31</session>
<nameOfHouse>衆議院</nameOfHouse>
<nameOfMeeting>内閣委員会</nameOfMeeting>
<issue>10号</issue>
<date>1959-02-24</date>
<speechOrder>25</speechOrder>
<speaker>齋藤正</speaker>
<speech>
○齋藤（正）政府委員　先ほど申し上げましたように、国語審議会としては、審議会で一応各方面の批判を仰ぐための中間的な報
```


図６　「検索用 API」での検索結果②

図６の３行目を見ると，「<startRecord>101</startRecord>」となっている。このことから，検索結果の101

番目からのXML文書（つまり，(9)＝図５の続き）が戻されていることがわかる。

繰り返し述べているように，(9)の検索URLによる「検索件数」は1,445件であるので，以下のように，15回分，検索URLを入力すれば，1947年５月１日から1960年12月31日までの国会において，「国語」を含む発言のすべてを入手することができる，というわけである。

(12) ①http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=1&maximumRecords=100&any=国語&from=1947-05-01&until=1960-12-31：
1件目から100件目の結果を戻す

②http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=101…（以降は(12)①と同様）：
101件目から200件目の結果を戻す

③http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=201…（以降は(12)①と同様）：
201件目から300件目の結果を戻す

④http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=301…（以降は(12)①と同様）：
301件目から400件目の結果を戻す

【中略】

⑬http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=1201…（以降は(12)①と同様）：
1,201件目から1,300件目の結果を戻す

⑭http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=1301…（以降は(12)①と同様）：
1,301件目から1,400件目の結果を戻す

⑮http://kokkai.ndl.go.jp/api/1.0/speech?startRecord=1401…（以降は(12)①と同様）：
1,401件目から1,445件目の結果を戻す

(12)の作業が終了すれば，今度は(9)の検索URLの「from」と「until」を適当な検索対象期間（「検索件数」が1,000件を大きく超えない範囲）で指定し，同様の作業を繰り返せばいい，ということになる。

このように，「検索用API」での「検索件数」が多い場合，単調な作業の繰り返しをしなければならない。既に繰り返し述べているように，「検索用API」において，１回の検索結果として表示されるXML文書の上限は，「発言単位」の検索で最大100件，「会議単位」の検索で最大５件である。したがって，５節でも述べたとおり，「検索件数」が多い場合は，全てのXML文書を入手するために同種の作業を繰り返さなければならない。このことをどのように解釈するかであるが，少なくとも，インターネットを使う程度のコンピュータリテラシーしか持ち合わせていない筆者にとっては，それほど大きな負担ではなかったということを付記しておく。実際，岡田（2018b）では，「国語」「日本語」あわせて数万件の検索結果を取り扱うことになったが，検索自体は１日で完了している。１度，基本となる検索URLを作成できれば，あとは上述した通り，検索URLの「startRecord」の数値と検索対象期間を指定しなおせばいいだけだからである。

なお，自分が検索したい語・表現が「国会会議録」にどの程度含まれているかは，「検索システム」を利用しておおよその目処を付けておくと良い。検索結果をダウンロードせずに，単純に「検索件数」を把握するだけであれば，「検索用API」よりも「検索システム」を利用するほうがずっと便利だからである。たとえば，「検索用API」を利用する際には，(9)に示したように，検索結果が1,000件を大幅に超えないように検索対象期間を設定する必要があるが，筆者は「検索用API」の検索URLを作成する前に，あらかじめ「検索システム」を利用して適当な検索対象期間の目処を付けたことを，参考までに記しておく。

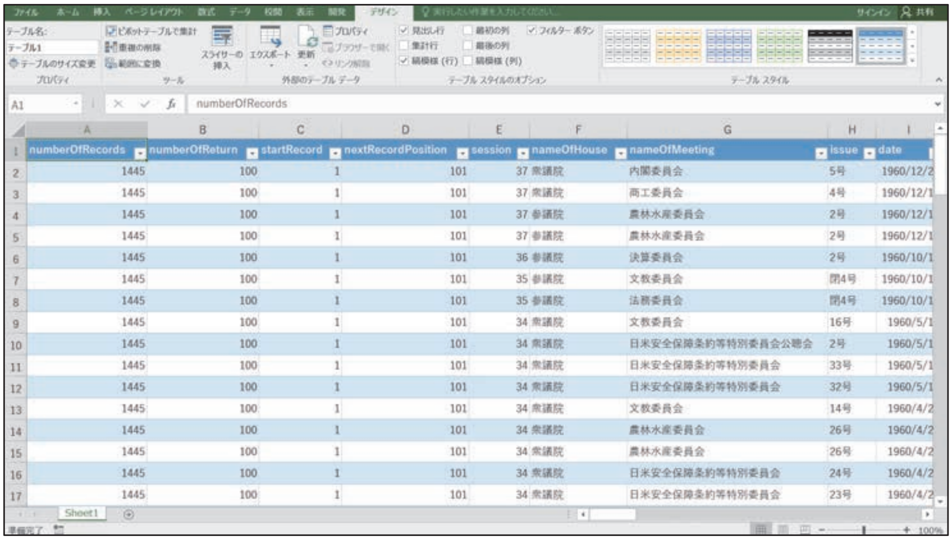
６．５　「検索用API」の検索結果の活用

前節での図５や図６で示したとおり，「検索用API」の検索結果は，XML文書で示される。そのXML文書をどのように活用するかであるが，岡田（2018b）においてはMicrosoft社の表計算ソフトExcelを利用して，「検索用API」の検索結果を分析した[16]（ExcelでXML文書をインポートする方法の説明は本稿での目的からは逸れるので詳述しないが，たとえば，Excel2016の場合，「開発」タブ→「インポート」→「XMLのインポート」で可能である）。

「検索用API」での検索の結果得られたXML文書をExcelで開くと，本節末尾に示した図７から図９のようになる。それぞれの列に記載されている情報は，以下の通りである。

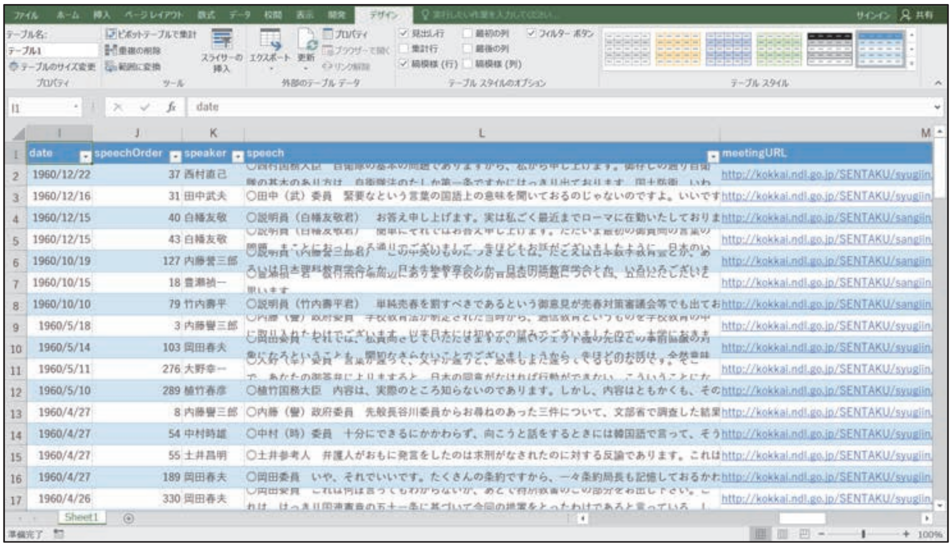
(13) A列：　総「検索件数」（図７の場合は1,445件）
B列：　XML文書として戻された件数（図７の場合は100件）

Ｃ列：　当該XML文書で戻された検索結果の開始通し番号（図７の場合は１番目）
Ｄ列：　当該XML文書で戻された検索結果に続く検索結果の通し番号（図７の場合は101番目）
Ｅ列：　当該発言がされた国会回次（図７の２行目の場合は第37回国会）
Ｆ列：　当該発言がされた院名（図７の２行目の場合は衆議院）
Ｇ列：　当該発言がされた委員会名（図７の２行目の場合は内閣委員会）
Ｈ列：　当該発言がされた委員会の号数（図７の２行目の場合は５号）
Ｉ列：　当該発言がされた年月日（図８の２行目の場合は1960年12月22日）
Ｊ列：　当該発言がされた会議における当該発言の通し番号（図８の２行目の場合は37）
Ｋ列：　当該発言の発言者（図８の２行目の場合は西村直己氏）
Ｌ列：　検索対象語を含む発言
Ｍ列：　当該検索結果に対する「国会会議録」（HTML文書）へのリンク
Ｎ列：　当該検索結果に対する「国会会議録」（PDF文書）へのリンク



	numberOfRecords	numberOfReturn	startRecord	nextRecordPosition	session	nameOfHouse	nameOfMeeting	issue	date
2	1445	100	1	101	37	衆議院	内閣委員会	5号	1960/12/2
3	1445	100	1	101	37	衆議院	商工委員会	4号	1960/12/1
4	1445	100	1	101	37	参議院	農林水産委員会	2号	1960/12/1
5	1445	100	1	101	37	参議院	農林水産委員会	2号	1960/12/1
6	1445	100	1	101	36	参議院	決算委員会	2号	1960/10/1
7	1445	100	1	101	35	参議院	文教委員会	閉4号	1960/10/1
8	1445	100	1	101	35	参議院	法務委員会	閉4号	1960/10/1
9	1445	100	1	101	34	衆議院	文教委員会	16号	1960/5/1
10	1445	100	1	101	34	衆議院	日米安全保障条約等特別委員会公聴会	2号	1960/5/1
11	1445	100	1	101	34	衆議院	日米安全保障条約等特別委員会	33号	1960/5/1
12	1445	100	1	101	34	衆議院	日米安全保障条約等特別委員会	32号	1960/5/1
13	1445	100	1	101	34	衆議院	文教委員会	14号	1960/4/2
14	1445	100	1	101	34	衆議院	農林水産委員会	26号	1960/4/2
15	1445	100	1	101	34	衆議院	農林水産委員会	26号	1960/4/2
16	1445	100	1	101	34	衆議院	日米安全保障条約等特別委員会	24号	1960/4/2
17	1445	100	1	101	34	衆議院	日米安全保障条約等特別委員会	23号	1960/4/2

図７　Excelにインポートした「検索用API」の検索結果（XML文書）①



	date	speechOrder	speaker	speech	meetingURL
2	1960/12/22	37	西村直己	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/syugiin
3	1960/12/16	31	田中武夫	○田中（武）委員　緊要なという言葉の国語上の意味を聞いておるのじゃないのですよ。いいです	http://kokkai.ndl.go.jp/SENTAKU/syugiin
4	1960/12/15	40	白幡友敬	○説明員（白幡友敬君）　お答え申し上げます。実は私ごく最近までローマに在勤いたしておりま	http://kokkai.ndl.go.jp/SENTAKU/sangiin
5	1960/12/15	43	白幡友敬	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/sangiin
6	1960/10/19	127	内藤誉三郎	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/sangiin
7	1960/10/15	18	豊瀬禎一	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/sangiin
8	1960/10/10	79	竹内壽平	○説明員（竹内壽平君）　単純売春を罰すべきであるという御意見が売春対策審議会等でも出てお	http://kokkai.ndl.go.jp/SENTAKU/sangiin
9	1960/5/18	3	内藤譽三郎	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/syugiin
10	1960/5/14	103	岡田春夫	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/syugiin
11	1960/5/11	276	大野幸一	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/syugiin
12	1960/5/10	289	植竹春彦	○植竹国務大臣　内容は、実際のところ知らないのであります。しかし、内容はともかくも、その	http://kokkai.ndl.go.jp/SENTAKU/syugiin
13	1960/4/27	8	内藤譽三郎	○内藤（譽）政府委員　先般長谷川委員からお尋ねのあった三件について、文部省で調査した結果	http://kokkai.ndl.go.jp/SENTAKU/syugiin
14	1960/4/27	54	中村時雄	○中村（時）委員　十分にできるにかかわらず、向こうと話をするときには韓国語で言って、そう	http://kokkai.ndl.go.jp/SENTAKU/syugiin
15	1960/4/27	55	土井昌明	○土井参考人　弁護人がおもに発言をしたのは求刑がなされたのに対する反論であります。これは	http://kokkai.ndl.go.jp/SENTAKU/syugiin
16	1960/4/27	189	岡田春夫	○岡田委員　いや、それでいいです。たくさんの条約ですから、一々条約局長も記憶しておるかわ	http://kokkai.ndl.go.jp/SENTAKU/syugiin
17	1960/4/26	330	岡田春夫	[illegible]	http://kokkai.ndl.go.jp/SENTAKU/syugiin

図８　Excelにインポートした「検索用API」の検索結果（XML文書）②

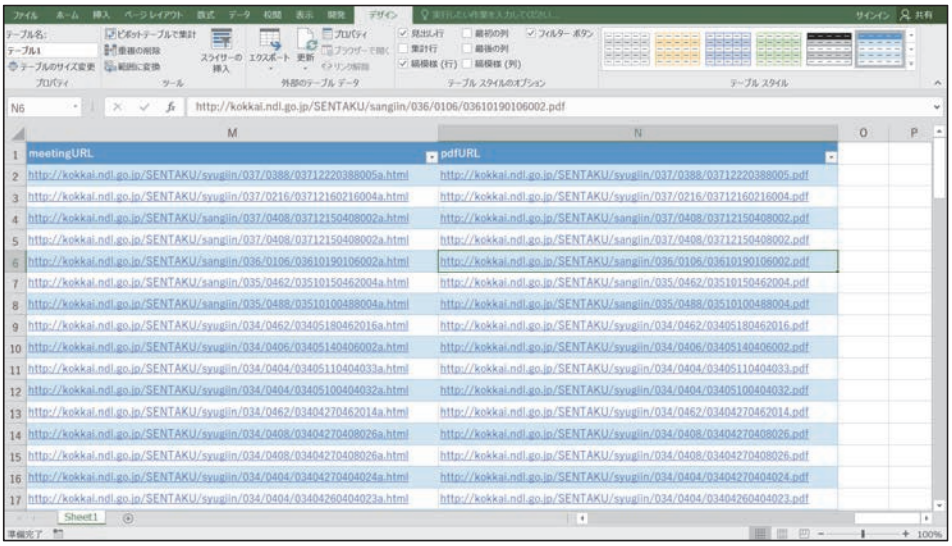


図９　Excel にインポートした「検索用 API」の検索結果（XML 文書）③

ある程度のコンピュータリテラシーを持ち合わせていないと，図５や図６に示したようなXML文書の形式示される「検索用API」の検索結果は活用しにくいと思われるが，図７から図９に示したようなエクセル形式であれば，（筆者のように）WordやExcelを使う程度の能力しか持ち合わせていない人間であっても，ある程度の分析，考察が可能となる。

最後になるが，図７から図９を見ると，「検索用API」での「発言単位」の「検索件数」は，（当然であるが）「発言単位」でカウントされていることが読み取れる。すなわち，「検索用API」での「検索件数」（図５や図６で，<numberOfRecords>として表示される件数）は，検索語が含む「発言単位」の数ということになる。したがって，たとえば一つの「発言単位」に100回，検索語が現れていたとしても，「検索用API」での「発言単位」の「検索件数」は「１」となる。「検索用API」での「発言単位」の「検索件数」＝用例数ではないことには注意したい（この点については，７節の議論も参照）

７．おわりに―「国会会議録」を利用した日本語研究，特に計量的研究の精緻化・深化を目指して―

本稿では，日本語研究資料としての「国会会議録」の性格を概観した上で，「国会会議録」を利用した従来の日本語研究で多用されてきた「検索システム」の問題点を指摘した。そのうえで，「国会会議録」を利用した日本語研究の新たな手法として「検索用API」を利用する選択肢を提案し，「検索用API」の利用方法について具体例をあげつつ概観してきた。本節では，本稿のまとめとして，「検索用API」をすれば，「国会会議録」を利用した日本語研究，特に計量的研究の精緻化・深化することが可能なのではないかと述べる。

３節でも述べたとおり，（「国会会議録」を利用した従来の日本語研究で多用されてきた）「検索システム」には，日本語研究に利用するにはいくつかの問題点が存在している。中でも，D「「検索件数」が検索対象語の用例数（出現度数）ではない」という問題点と，E「検索の結果には，検索対象語の一部分しか一致しないものもヒットしてしまう」という問題点は，特に計量的な研究を行う際には，場合によっては致命的な影響を与えてしまう可能性もある。

この問題点に対する「検索システム」を利用した従来の日本語研究の対応策（の一部）については，既に３節で紹介したのでここでは繰り返さないが（茂木2008の試み），殊「国会会議録」を利用して計量的な研究を行う際には，本稿で利用することを提案した「検索用API」を活用することを強く推奨する事例を紹介したい。

筆者は，近年，言語政策（国語政策）が検討される場の一つである国会において，種々の言語問題がどの

ように議論されてきたのか関心を持ち，言語政策（国語政策）に関連するいくつかの語について，「検索システム」や「検索用API」を利用し検索を行っている（岡田2017・2018a・2018b）。そのいくつかの例をまとめたものが，以下の表２である。

表２　「検索システム」と「検索用 API」での検索結果

検索語	「検索システム」での検索件数	「検索用API」での検索結果		
		検索件数	目視で確認した用例数（複合語・派生語用法も含む）	目視で確認した用例数（単独用法のみ）
「多言語」	213	291	356	146
「手話」	602	1,589	3,746	1,465
「方言」	1,016	1,105	429	425

表２の凡例は，以下の通りである。

Ⅰ）「検索システム」での検索件数：
　「検索語」を「検索システム」で検索した結果，表示される検索件数

Ⅱ）「検索用API」での検索結果
　ⅰ）検索件数；
　　「検索語」を「検索用API」で検索した結果，表示される検索件数（図５，図６では，<numberOfRecords>として表示される件数）
　ⅱ）目視で確認した用例数[17]（複合語・派生語的用法も含む）：
　　「検索用API」の検索結果，得られるXML文書を利用し，検索結果を目視で確認し，検索結果から「検索語」の部分一致を除外した件数
　ⅲ）目視で確認した用例数（複合語・派生語的用法も含む）：
　　Ⅱ－ⅱの結果から，複合語・派生語的用法を除いた件数

なお，Ⅱ－ⅱとⅡ－ⅲの違いを「手話」を例に説明すると，Ⅱ－ⅰには「手話」「日本手話」「標準手話」といった用例が含まれるが，Ⅱ－ⅱには「手話」という用例しか含まれない，ということである。

最後に，表２の数値を求めるにあたって種々の作業を行なった結果，「検索システム」の「検索件数」と「検索用API」の「検索件数」が大きく食い違う場合があるのは勿論のこと，「検索用API」の検索結果の中にもいわゆる「検索語」の「部分一致」の例が相当数含まれる場合があることも判明したことを報告しておきたい。特に，「方言」については，筆者はdialectという意味での「方言」を検索したかったのであるが，「検索用API」の結果を見ると，「両方言った」「双方言った」「皆さん方言った」という，筆者が当初想定もしていなかったものが多数ヒットし，喫驚した。このことからも，「国会会議録」の検索結果については，件数が多くとも，発言内容を自身の目で確認し，吟味する必要があると，改めて痛感した次第である。

もっとも，この問題点は，３節でも触れたとおり，既に，「検索システム」を利用した研究の時点で指摘されていたものである（茂木2008，松田2012など）。ただ，「検索システム」の仕様上，「検索システム」の検索結果を簡単にダウンロードできないため，（註７でも触れたとおり）「「検索件数」からおおまかな使用傾向を捉えることはできるかもしれないが，量的な議論を行う際には注意する必要がある」（茂木2008）という警鐘は鳴らされていたものの，問題解決のための具体的な方策が見出されていない状態だったと思われる。しかし，今回，筆者が紹介した「検索用API」を利用すれば，この種の問題が完全に解決されるわけではないが，少なくとも，従来の方法よりは随分負担の少ない形で，「国会会議録」を利用した計量的な日本語研究の実践が可能になるのではないかと考える。そのような意図を込めて，本稿のタイトルを「日本語コーパスとしての「国会会議録検索システム検索用API」―計量的研究の精緻化・深化の可能性―」としたのであるが，今後，筆者以外にも，「検索用API」を利用し，現代日本語の計量的な研究を実践なさる方が現れ，「国会会議録」を利用した日本語研究が精緻化，深化することを願ってやまない[18]。

謝辞

本稿は，「国会会議録」を利用した日本語研究の様々な試み，先行研究の延長線上に位置します。中でも，「国会会議録」を利用した日本語研究の可能性を提示され，実践を積み重ねられてきた松田謙次郎先生（松蔭女

子学院大学教授）の様々なご論考には，大変刺激を受けました。特に，２節，３節については，松田先生の先駆的なご研究がなければ，執筆できませんでした。常に先駆的な試みをされ，筆者に日々，新しい視点をくださる松田先生に，深く御礼申し上げます。

また,「国会会議録」の作成に関わられた／関わっていらっしゃるみなさま, さらには「検索システム」と「検索用API」の開発に当たられた／あたっていらっしゃるみなさまのお力がなければ，そもそも，本稿は成立しえませんでした。関係者のみなさまに，この場をお借りして，心からの敬意と謝意をお伝えしたいと存じます。

さらに，７節での議論の用例の整理に使用させていただいた「日本語KWIC索引生成ソフトウェアKWIC」の製作者でいらっしゃる田野村忠温先生（大阪大学教授）にも, 心より御礼申し上げます。と同時に, 田野村先生が当該ソフトウェアのマニュアル（http://www.tanomura.com/research/KWIC/files/KWIC.pdf）の冒頭でお書きになった以下のお言葉は，筆者のようには「安易な」研究を行っている者は，重く受け止めなければいけないと痛感した次第です。

> コーパス利用の普及のためには簡単に使えるコーパス処理ソフトウェアの存在が不可欠だと思われるが、残念ながら日本語に関しては研究者が手持ちの電子テキストをそのまま使って簡便にKWIC索引を生成することのできるソフトウェアがなかった。
>
> そうしたことから、2007年に（中略）Windows上で作動する日本語KWIC索引生成ソフトウェアを試作してみた。初めて使うRubyによって短期間で書いた素朴なもので、遠からず高機能で使いやすいKWICソフトウェアが作られて普及するものと予想していたが、作成から10年以上経つ今も多くの方に使われている。（中略）これは作った甲斐があったという意味では喜ばしいが、コーパス日本語研究の水準の反映と受け止めれば寂しい現実である。

なお, 本稿の内容の一部は, 日本語学会2018年度春季大会（2018年５月20日･明治大学駿河台キャンパス）で行ったブース発表（岡田2018b）に基づきます。学会当日，様々なご意見をくださったみなさまに，心より御礼申し上げます。勿論，本稿の一切の責の一切は，筆者に帰します。

付記

本文中にあるURLは，特に本文中でことわりをしない限り，2015年６月25日に最終確認を行った。

註

1)「国立国語研究所が作成・提供する，日本語学・日本語教育に関係する学術論文情報検索のためのリファレンスデータベース」（中野真樹･渡辺由貴2013）である「日本語研究･日本語教育文献データベース」（https://bibdb.ninjal.ac.jp/bunken/）で,「国会会議録」を検索語に指定して検索すると，松田（2004）以前にも「国会会議録」に関する文献が数件ヒットする。しかし，それらの文献は，衆議院記録部･参議院記録部による『国会会議録用字例』や『国会会議録用語集』，国会会議録と速記の関係を論じたもの，国会会議録に対する検索技術を論じたものである。それゆえ，松田（2004）は「国会会議録」を日本語研究の資料として利用する可能性を論じた先駆的な文献と位置づけて良いと思われる。

2) 註１でも利用した「日本語研究・日本語教育文献データベース」で,「国会会議録」を検索語に指定して検索すると，2018年６月25日現在，57件の文献がヒットする（そのうち，松田2004以降に発表された文献は50件である）。なお，当該データベースに格納されている総データ件数は，2018年６月現在で約233,000件とのことである（https://bibdb.ninjal.ac.jp/bunken/index.php?mode=about）。

3) 松田（2004，2008）によると,「国会法と議院規則によって規定されている不穏当発言の例は，無礼な発言または他人の私生活に関する発言，敬称の不使用，書籍等の朗読，私語，妨害的発言，議題外の発言である」とのことである。

4) プレス･コードに抵触する発言が「国会会議録」に反映されないということは,「ある種の語彙が「不適当」なものとしてシステマティックに削除されている」場合も考えられ，特にそのようなケースに該当する語彙を研究する際には障害になるであろうと，松田（2004，2008）は指摘している。

5) 松田謙次郎ほか（2008）では，国会の動画音声と「国会会議録」の記録を対照させ,「国会会議録」での

どのような整文が施されているのか，その一端を明らかにしている。

6) もっとも，国会審議の模様は，オンライン上で動画・音声を確認することができる。URLは以下の通り。

・「衆議院　インターネット審議中継」：　http://www.shugiintv.go.jp/jp/index.php
第174回国会（2010年1月18日）以降の審議を視聴可能

・「参議院　インターネット審議中継」：　http://www.webtv.sangiin.go.jp/webtv/index.php
会期終了日から1年が経過した日まで視聴可能

「国会会議録」と上記のサイトで視聴可能な動画・音声とを併用すれば，「国会会議録」での記録内容が実際の発言に忠実か否かを確認できるだけはなく，「国会会議録」だけを利用した研究では不可能な音声，あるいは身振り手振りといった非言語情報までを視野に入れた研究も可能である。

なお，政策研究大学院大学「比較議会情報プロジェクト」（代表者：　増山幹高氏）によって構築された「国会審議映像検索システム」（http://gclip1.grips.ac.jp/video/。2012年12月一般公開）を利用すれば，「国会図書館の提供する会議録と衆参両院の事務局が配信する審議動画をリンクさせ，発言のキーワード検索から審議映像をピンポイントで再生すること」が可能になった。この「国会審議映像検索システム」を日本語研究に利用した試みも存在する（松田謙次郎2016）。

7) 茂木（2008）は，「アイドリングストップ」「アメニティ」「マニフェスト」という3語について，「検索システム」の「検索件数」と実際の用例数を求めたところ，以下に示すように，「検索件数」と実際の用例数の関係は様々であることを指摘し，「「検索件数」からおおまかな使用傾向を捉えることはできるかもしれないが，量的な議論を行う際には注意する必要がある」と注意を喚起している。

	検索件数	用例数
「アイドリングストップ」	27	48
「アメニティ」	278	560
「マニフェスト」	326	1,954

8) 松田（2012）は，入門書にふさわしい軽妙洒脱な表現で，以下のように述べている。

悲しいかな，このためには手作業ですべての会議を当たるしかありません。検索件数からわかるのはあくまで会議数だけで，最低限これだけの回数は使われているということしかわかりません。これをクリックすればすべてを数えてくれるような「秘密のボタン」などもないようです。

9) 1回に最大5件のXML文書しか取得できない「会議単位」の検索の場合は，負担はほとんど減らないのではないかという意見もあろう。確かにそうなのであるが，「国会会議録」を利用した日本語研究の主な研究，すなわち語彙や文法の研究においては，「発言単位」の検索で十分対応できる。つまり，「国会会議録」を日本語研究に利用する場合は「検索用API」の「発言単位」を活用することで十分であり，それゆえ「検索用API」を活用した日本語研究は「検索システム」以上の利便性を享受できるのではないかと筆者は考えている。

10)「特定の主題や政策キーワードに着目して会議録のデータを取得することにより，ある主題の扱いが時系列でどのように変化したか，国会でどのような議論が行われたか等について，計量的な分析を行う」研究のこと（川瀬・清水2015）。

11) このような情報公開は，国会での審議に関する情報公開は，「国会会議録」以外に，註6でも紹介したような動画配信によっても行われている。ただ，従来は，「国会会議録」と国会審議の動画配信は独立した形で行われている。註6で紹介した「比較議会情報プロジェクト」は，別個の形で公開されている「国会会議録」の国会審議の動画配信を結びつける試みであるが，川瀬・清水（2015）は，「各システムが個別にAPIを提供するようになれば，このような（引用者註：　「比較議会情報プロジェクト」のようなマッシュアップの取組は容易になり，制度や組織を超えた活用が可能となる」と述べている。

12) 川瀬・清水（2015）では，「国会でどのような議論がどれだけ行われているのか，誰がどの程度活動しているのかという情報は，冊子体で公開された会議録を見ているだけでは十分に把握しにくい」が，「本APIを使うことで，会議・発言の量や回数などを量的に把握することが容易になる」と述べられている。

13)「検索用API」を利用した一連の筆者の試み（岡田祥平2017，2018a，2018b）は，筆者の専門分野，関心から，「検索用API」を「日本語言語コーパスとしての活用」することを模索するものであったつもりであるが，実質的に行ったことは「特定の主題や政策研究における活用」する試み，あるいは「国会活動の量的

な可視化」することの試み，模索であったかもしれない。

14) 川瀬･清水（2015）では，この部分で，松田［編］（2008）と山本（2008）を参考文献としてあげている。

15) なお,「検索用 API」のホームページには, どの要素に URL エンコードを施すが必要か, 明記されていない。以下の(2)に示す諸要素は, 筆者が「検索用 API」の検索をする中で, 経験的に得た知見である。したがって, 漏れなどがある可能性があることを，あらかじめお断りしておく。

16) XML 文書を Excel 上で利用するのは効率の良いこととも思えないのであるが，筆者のようにコンピュータリテラシーが高くない人間にとっては作業の効率化に一定以上の意味はあったことを申し添えておきたい。

17) 目視による用例の確認の際には，田野村忠温氏（大阪大学教授）による「日本語 KWIC 索引生成ソフトウェア KWIC」（http://www.tanomura.com/research/KWIC/）を使用させていただいた。田野村氏によるこのソフトウェアは，検索文字列の先行文脈を基準にソートした結果と後続文脈を基準にソートした結果の２種類が一つのエクセルファイルとなって出力されるなど機能があり, 本稿での用例の整理に大きな助けとなった。

18) 本稿で紹介した方法は，インターネットと Microsoft 社の製品しか使えないという程度のコンピュータリテラシーが持ち合わせていない筆者でも，実践できるものである。したがって，日本語研究に関心のある学部生であっても，少々訓練を積めば，筆者が紹介した以上の検索や研究が可能になるはずである。

参考文献

井上史雄（1983）「ジュニア言語学　気づかない方言」『言語』第 12 巻６号

岡田祥平（2017）「“日本における「多言語」”を国会はどのように論じてきたのか―「国会会議録検索システム」を利用した経年的な考察の試み―」新潟県ことばの会・平成 29 年度研究集会発表資料

岡田祥平（2018a）「国会では「手話」がどのように論じられてきたのか－「国会会議録検索システム検索用 API」を利用した経年的な考察－」『社会言語科学会第 41 回大会発表論文集』

岡田祥平（2018b）「国会では「国語」「日本語」という語がどのように使用されてきたのか―「国会会議録検索システム検索用 API」を利用した日本語研究の一実践の紹介を兼ねて―」『日本語学会 2018 年度春季大会予稿集』

沖　裕子（1992）「気づかれにくい方言」『言語』第 21 巻 11 号

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 巻 12 号

中野真樹・渡辺由貴「国立国語研究所「日本語研究・日本語教育文献データベース」の有用性」

樋口耕一（2014）『社会調査のための計量テキスト分析　内容分析の継承と発展を目指して』ナカニシヤ出版

松田謙次郎（2004）「言語資料としての国会会議録検索システム」『TALKS: Theoretical and Applied Linguistics at Kobe Shoin』No.7

松田謙次郎（2008）「国会会議録検索システム総論」松田謙次郎［編］『国会会議録を使った日本語研究』ひつじ書房

松田謙次郎（2012）「国会会議録をつかう」日比谷潤子［編著］『はじめて学ぶ社会言語学―ことばのバリエーションを考える 14 章―』ミネルヴァ書房

松田謙次郎（2016）「国会審議映像検索システムの社会言語学的応用について」『社会言語科学会第 38 回大会発表論文集』

松田謙次郎［編］（2008）『国会会議録を使った日本語研究』ひつじ書房

松田謙次郎・薄井良子・南部智史・岡田裕子（2008）「国会会議録はどれほど発言に忠実か？―世文の実態を探る―」松田謙次郎［編］『国会会議録を使った日本語研究』ひつじ書房

丸山岳彦･柏野和佳子（2014）「サンプリング」前川喜久雄［監修］･山崎　誠［編］『講座日本語コーパス　２．書き言葉コーパス―設計と構築―』朝倉書店

茂木俊伸（2008）「国会会議録における行政分野の外来語」松田謙次郎［編］『国会会議録を使った日本語研究』ひつじ書房

山口昌也（2013）「付録B　コーパス検索ツール（２）　全文検索システム『ひまわり』」前川喜久雄［監修・編］『講座日本語コーパス　１．コーパス入門』朝倉書店
山本和英（2008）「自然言語処理での国会会議録の利用」松田謙次郎［編］『国会会議録を使った日本語研究』ひつじ書房