E-009

品詞のパターンマッチングを用いた帝国議会議事速記録における可能表現の自動抽出
Automatic Extraction of Possible Expressions in the Stenographic Record of the Imperial Diet using Part-of-Speech Pattern Matching

北野 勇太[1)]　横山 想一郎[2)]　山下 倫央[2)]　川村 秀憲[2)]　伊藤 孝行[3)]
Kitano Yuta　Yokoyama Soichiro　Yamashita Tomohisa　Kawamura Hidenori　Ito Takayuki

1　はじめに

1.1　研究背景

帝国議会議事速記録は，近代日本における国会の議事録である．57 年にわたって蓄積されているため，分量が非常に多い．さらに口語での対話が記録されているため，当時の人々がどのような言葉を話していたのかを研究する重要な資料となる．帝国議会議事速記録の分析は，近代の議会の場での言語表現の使用状況を明らかにすることができ，学術的価値が高い．その分析例の 1 つに，可能表現がある．

可能表現は，物事の出来る出来ないを表す日本語表現のことである．これは議会という場において議員が頻繁に使用していた表現で，議論に大きな影響を及ぼしていたと推察される．

しかし，帝国議会議事速記録の多くがテキストデータ化されておらず，画像データとして保存されている．そのため帝国議会議事速記録の分析は，速記録全体を対象とするような大規模な分析は困難であった．これは光学文字認識（OCR）技術 [1] により解決に向かっている．本研究では，それに先駆け，帝国議会議事速記録に着目し調査を行う．

1.2　研究目的

本研究では，帝国議会議事速記録における議員の可能表現の使用状況が時間とともにどのように変遷してきたかを明らかにするための前段階である，可能表現を自動で抽出する手法の提案を行う．また，会議の開催時期による自動抽出への影響を調査する．

1.3　関連研究

神田は小説を対象として現代東京語における可能表現の変遷を分析した [2]．伊藤らは帝国議会議事速記録を対象として東北地方出身議員の使用する可能表現について分析した [3]．同一の意味を持つ日本語表現の自動抽出としては，Web 上に書き込まれたユーザーの商品やサービスに対する評価情報を収集するという目的で，評価表現の抽出に関する研究がよく行われる．評価表現の自動抽出手法として，小林らは，意見に関わる表現を文型や評価対象・評価値属性に基づいてフィルタリングを行う手法を提案している [4]．また高野らは，評価表現辞書と構文解析を用いて因果関係から評価表現を抽出する手法を提案している [5]．本研究では伊藤ら [3] の研究に基づいて，帝国議事会議速記録を対象とした可能表現の自動抽出に取り組む．

1）　北海道大学大学院情報科学院
2）　北海道大学大学院情報科学研究院
3）　北海道大学大学院メディア・コミュニケーション研究院

2　帝国議会議事速記録

帝国議会議事速記録は，1890 年から 1947 年までの 57 年間開かれていた帝国議会の中で話されていた内容が，口語で記録されたものである．一般的な議事録は議論の要点のみを記録するが，帝国議会議事速記録は話されていた内容がそのまま記録されているという点が異なる [6]．口語で記録された資料は当時に話されていた言葉や文法を分析する上で有利な要素であり，帝国議会議事速記録は分量も多いことから，近代の議会における言語表現の分析を行う上では質の高い資料といえる．

帝国議会の会議は本会議と秘密会の 2 つに分けられる．そのうち本会議の速記録は帝国議会会議録検索システム [7] という Web ページで公開されている．ここでは，1890 年 11 月から 1945 年 8 月までの戦前にあたる期間は 24,894 件が画像データのみで，1945 年 8 月から 1947 年 3 月までの戦後期間は 1,249 件が画像データとテキストデータで公開されている．図 1 は会議録本文の一例である．

大量の文書データの 95%以上が画像データ形式での公開であり，この会議録の分析には事前にテキスト化を行うことが必須である．このテキスト化しかし，光学文字認識（OCR）技術 [1] によりテキスト化の問題は解決に向かいつつある．これにより，帝国議会議事速記録全体

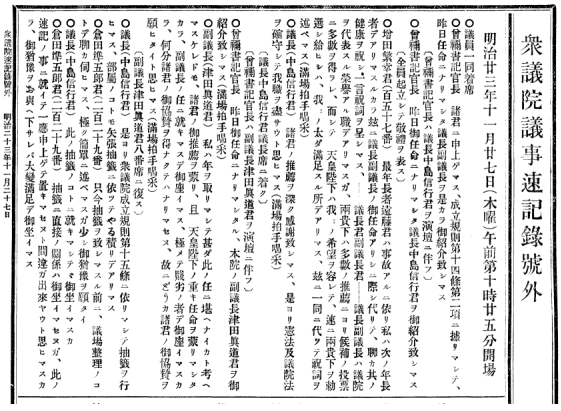
衆議院議事速記録號外
明治廿三年十一月廿七日（木曜）午前第十時廿五分開場

図 1　第 1 回帝国議会 衆議院 奉答文に関する件（明治 23 年 11 月 29 日）

を対象とした大規模な分析を行えるようになる見込みが出てきているというのが現在の状況である．一方，分析の前段階となる対象日本語表現の抽出手法はまだ確立されていない．

3 可能表現

3.1 可能表現の形

可能の意味を表す日本語表現は多数存在する．これらの自動抽出を行うには，この表現形式をいくつかに区分する必要がある．本研究では，帝国議会議事速記録の可能表現の分析例 [3] における可能表現の区分を利用した．区分は以下に示す 9 つである．

1. 連体形＋コト＋助詞＋「出来ル」
2. 名詞＋助詞＋「出来ル」
3. 名詞＋「出来ル」
4. 得ル
5. レル・ラレル
6. 可能動詞
7. 能フ
8. ナラヌ
9. 重複型

以下では，各区分の用例や特徴について述べる．

3.1.1 連体形＋コト＋助詞＋「出来ル」

動詞の連体形に，「〜ことが出来る」「〜ことも出来る」などが付いた形である．以下に用例を示す．

- 「走ることが出来る。」
- 「走ることが出来ないし、歩くことも出来ない。」

3.1.2 名詞＋助詞＋「出来ル」

名詞に，「〜が出来る」「〜は出来る」などが付いた形である．以下に用例を示す．

- 「全力疾走が出来る。」

3.1.3 名詞＋「出来ル」

3.1.2 節の形に似ているが，名詞と「出来る」の間に助詞がないという点が異なる．以下に用例を示す．

- 「全力疾走出来る。」

3.1.4 得ル

動詞の連用形に「得る」が付く形や，「〜すること」＋助詞＋「得る」の形などがある．これが当時「うる」と読まれていたのか「える」と読まれていたのかは不明である [3]．また，「手に入れる」という意味の動詞「得る」とは表記が同じであるが，こちらは可能の意味を持たないので区別が必要である．以下に用例を示す．

- 「考え得る限りの案を出す。」
- 「様々なことが起こり得る。」

3.1.5 レル・ラレル

助詞「れる」や「られる」には受身・可能・自発・尊敬の 4 つの意味がある [8]．これらの意味の判別は文脈情報が必要となる [9]．文脈情報が不足していると，意味を断定できない場合もあるので注意が必要である．ここでは可能の意味で使われる例を以下に示す．

- 「この料理は十分食べられる量だ。」
- 「この服は小さすぎて着られない。」

3.1.6 可能動詞

可能動詞とは，五段活用動詞を下一段活用に変化させた動詞である．「読める」「遊べる」などがこれに当たる．終止形が「エ段＋る」の形になることが 1 つの特徴である [10]．なお、「来れる」といったいわゆる「ら抜き言葉」とは明確に違うものである．「来れる」のもととなる動詞「来る」はカ行変格活用であり，五段活用動詞ではないためである．以下に用例を示す．

- 「この距離なら走れる。」
- 「私は泳げない。」

3.1.7 能フ

「能ふ」は，可能の意味を表すハ行四段活用動詞である．動詞の連体形の後ろに付くのが一般的である．以下に用例を示す．

- 「了解し能はざるところである。」
- 「多少の変動あることは能はない訳であります」

3.1.8 ナラヌ

「ならぬ」は「出来る」が可能の意味で使われる前に使われていた表現であり，禁止と可能の意味が存在する [11]．可能の意味で使われる際には「-ことはならぬ」「-ことがならぬ」といった形を取ることが多い．以下に用例を示す．

- 「公務のために欠席することはならぬ。」

3.1.9 重複型

重複型は，複数の可能表現が重複して出来た表現である．帝国議会議事速記録においては，同時期の言語資料に比べ重複型が多数出現する [3]．以下に用例を示す．

- 「資金の調達が出来得るのである。」
- 「議長限で出来得べきことでない。」

3.2 その他の可能表現

3.1.1 節から 3.1.9 節までの区分の他に，すでに可能表現としてわかっているものを挙げる．これらは議会中に複数回出現する表現で「出来る」「出来ない」の意味を持つため，可能表現として抽出されるべき表現である．

- 「出来るだけ」
- 「〜し兼ねる」
- 「可能」「不可能」

4　可能表現の自動抽出手法

日本語表現を抽出する方法として，先行研究の CRF を用いた手法や大規模言語モデルによる手法が考えられる．本研究では，3.1.1 節から 3.1.9 節までで述べた可能表現の区分，および 3.2 節で述べた可能表現に対応するように品詞の組み合わせルールを設定した．これらに合致するものを可能表現として抽出する．日本語は多様な表現が可能であるため，可能表現すべての抽出は困難であり，適合率と再現率はトレードオフとなる．抽出手順は以下の通りである．

1. 抽出対象の文書に対し，旧仮名口語 unidic を導入した fugashi による形態素解析を適用する．
2. ルールに合致する品詞の並びがある場合，その部分を可能表現として抽出する．

例えば「現状におきましてはとうてい遺憾ながら期待できぬのではないかと思うのであります。」という文章においては，「期待でき」の部分が「名詞＋出来る」（3.1.3 節）の形であるため，ここを可能表現として抽出する．

4.1　品詞の組み合わせルール

1. 動詞＋「こと」＋助詞 1 個以上＋副詞 0 個以上＋「出来る」
2. 名詞＋助詞 1 個以上＋副詞 0 個以上＋「出来る」
3. 名詞＋「出来る」
4. 五段活用動詞かつ，終止形が「エ段＋る」で終わる動詞
5. 動詞「能ふ」
6. 「こと」＋助詞＋「ならぬ」
7. 動詞＋「得る」
8. 「兼ねる」
9. 「可能」
10. 「出来るだけ」

3.1.5 節で述べたように，助動詞「れる」「られる」には複数の意味が存在し，その意味を判別するには文脈情報が必要である．そのため品詞の組み合わせによる抽出は難しく，本研究の提案手法では抽出対象から除外した．

5　実験

5.1　実験目的

4 章にて述べた手法を用いて，可能表現の自動抽出を試みる．本研究では，以下の 2 点を明らかにすることを目的とする．

- 品詞の組み合わせにより可能表現を抽出出来るか
- 会議時期の違いにより抽出精度にどの程度影響が生まれるか

表 1　対象会議の文字数と可能表現数

会議名	文字数	可能表現数
第 45 回帝国議会衆議院予算委員第四分科会陸軍の軍備に付て（秘密会）	28,015	87
第 91 回帝国議会 衆議院 本会議 第 2 号 昭和 21 年 11 月 27 日	28,363	46

5.2　実験設定

事前調査で可能表現の数が比較的多いと判明している，時期の違う「第 45 回会議秘密会」と「第 91 回本会議第 2 号」を選定した．第 45 回会議は 1922 年，第 91 回会議は 1946 年に開かれた会議である．この 2 会議の発言中に含まれる可能表現を、事前にすべて日本語学の研究者に依頼し人手で抽出した．これを正解データとして実験を行う．各会議における文字数と抽出した可能表現数は，表 1 に示した．ただし，2 章で述べたように，戦前の会議録はすべて画像データとして公開されておりテキストデータを入手することが難しい．そのため，すでにテキスト化されている秘密会のデータを用いた．本会議と秘密会で会議録の構成や発言者に大きな違いはないため，これらによる結果への影響は小さいと考えられる．提案手法の評価は適合率，再現率，および F1 スコアで行う．

5.3　結果

第 45 回会議と第 91 回会議に対して同じルールを適用すると，第 91 回側の自動抽出数が 15 と明らかに少なくなった．これは漢字や送り仮名といった表記の違いによるものであることが明らかになった．第 45 回会議と第 91 回会議で見られた主な表記の違いを表 2 に示す．

表 2　第 45 回会議と第 91 回会議における表記の違い

第 45 回の表記	第 91 回の表記
出来る	できる, 出来る
能はず	能わず
兼ねる	かねる

これらを修正した上で 4.1 節のルールを用いた結果を表 3 および表 4 に示した．どちらの会議においても，82%程度の可能表現を抽出することが出来た．「ならぬ」に関する可能表現は，正解データに存在しているが抽出数が 0 となっていた．これは「それがならぬ」のように前までの内容を踏まえて述べられている形だったためである．

5.4　考察

5.3 節で述べたような文脈情報が必要と考えられる表現を可能表現として抽出するには，品詞の組み合わせによる手法とは別のアプローチを取る必要があると考えられる．

また，表 2 のような表記の違いが議会中期と後期で現れたため，議会初期にもこのような違いがある可能性があると考えられる．特に仮名遣いの変化は形態素解析の

表 3　各会議におけるルールごとの抽出数と適合率

区分	第 45 回秘密会		第 91 回本会議第 2 号	
	抽出数	適合率	抽出数	適合率
1. 動詞＋「こと」＋助詞 1 個以上＋副詞 0 個以上＋「出来る」	18	0.833	8	1.000
2. 名詞＋助詞 1 個以上＋副詞 0 個以上＋「出来る」	62	0.774	23	0.957
3. 名詞＋「出来る」	2	1.000	3	1.000
4. 五段活用動詞かつ，終止形が「エ段＋る」で終わる動詞	19	0.316	7	0.000
5. 動詞「能ふ」	0	-	0	-
6. 「こと」＋助詞＋「ならぬ」	0	-	0	-
7. 動詞＋「得る」	16	0.938	8	0.750
8. 「兼ねる」	4	0.750	1	1.000
9. 「可能」	1	1.000	2	1.000
10. 「出来るだけ」	1	1.000	9	1.000

表 4　各会議における可能表現自動抽出の結果

会議	可能表現数	総抽出数	適合率	再現率	F1 スコア
第 45 回	87	107	0.673	0.828	0.742
第 91 回	46	55	0.691	0.826	0.752

関係で結果に大きな影響を及ぼすと推測されるため，より調査を進める必要がある．

さらに，第 45 回会議のテキストデータは現代の漢字を用いたデータであったが，第 91 回会議の会議録中には「將來」「戰爭」「從來」といった，現在の JIS 第 2 水準にあたる漢字が散見された．これらは現代日本では使用頻度の低い漢字であり，この漢字が可能表現に含まれていると形態素解析の辞書に登録されていないなどの理由で抽出が困難になる可能性がある．

そのため，可能表現の自動抽出の実現には，会議の開催時期による漢字や仮名遣いの違いへの対応が今後の課題として挙げられる．

6　おわりに

本研究では，帝国議会議事速記録から品詞の組み合わせによる可能表現の抽出を試みた．先行研究による可能表現の区分からルールを作成し，およそ 82%の可能表現を抽出することが出来た．

一方，可能表現区分にない日本語表現の登場や，現代との漢字表記の違いが見られた．本稿では第 45 回会議と第 91 回会議の文書を用い，可能表現を自動抽出する上で議会中期と後期間にどのような違いがあるかが明らかとなった．可能表現の自動抽出を実現するためには，議会初期についても調査が必要になる．

また、文脈情報が必要なため品詞の組み合わせによる抽出が難しいと述べた助動詞「れる」「られる」は数多く出現するため，これらの抽出も今後の課題として挙げられる．

参考文献

[1] 令和 3 年度 OCR 処理プログラム研究開発. https://lab.ndl.go.jp/data_set/ocr/r3_software/ 2023 年 6 月 10 日参照.
[2] 神田寿美子：現代東京語の可能表現について，日本文學，Vol.16，pp.70-84 (1961).
[3] 伊藤孝行ほか：『帝国議会衆議院議事速記録』 に於ける東北地方出身議員の可能表現について (2002).
[4] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一：意見抽出のための評価表現の収集，自然言語処理，Vol.12, No.3，pp.203–222 (2005).
[5] 高野敦子，池奥渉太，北村泰彦：因果関係に着目した口コミ Web サイトからの評価表現抽出，人工知能学会論文誌，Vol.24, No.3，pp.322–332 (2009).
[6] 国立国会図書館：日本-議会資料（帝国議会）. https://rnavi.ndl.go.jp/jp/politics/Japan-teikoku.html 2023 年 6 月 10 日参照.
[7] 帝国議会会議録検索システム. https://teikokugikai-i.ndl.go.jp/#/ 2023 年 6 月 10 日参照.
[8] 小矢野哲夫：現代日本語可能表現の意味と用法 (I)，大阪外国語大学学報，Vol.45，pp.83–98 (1979).
[9] 嶋真也，山本和英：助動詞「れる」「られる」の自動意味判別，言語処理学会 第 25 回年次大会 発表論文集.
[10] 山岡政紀ほか：可能動詞の語彙と文法的特徴，日本語日本文学，No.13，pp.A1–A36 (2003).
[11] 渋谷勝己：日本語可能表現の諸相と発展，大阪大學文學部紀要，Vol.33，No.1，pp.i–262 (1993).