イ形容詞文における丁寧語使用の歴史的変化 *
状態空間モデルを用いた時系列分析

山田彬尭

1. はじめに

規範文法におけるイ形容詞文に使用される丁寧語の取り扱いは、20世紀後半に大きな転換を迎えている。1916年から2009年までに出版された16の規範文法書や明治期から昭和期の国定教科書を調査した川口(2014)は、(i) 今では人口に膾炙した(1)a のような「イ形容詞＋です」という表現が 20 世紀後半に至るまでは規範的だとは見なされてこなかったことを明らかにしており、 (ii) 代わりに(1)b のような「(ウ音便の)イ形容詞＋ございます」が、意図した意味を表現するために推奨されていたことを指摘している。

(1) a. 富士山は 美しいです。 新形式
b. 富士山は 美しゅうございます。 旧形式

この川口(2014)の研究は、丁寧語の構文の確率的交替全般に関わる、通時的、共時的なバリエーションについて大局的な傾向を詳らかにした点で大きな意義を持つ。ただし、下記の項目については十分な検討が尽くされたとは言い難い。

(2) 先行研究において取り残されているリサーチクエスチョン
a. 新形式はいつごろから用いられてきたのか。
b. どのような社会言語学的環境で新形式はその使用を拡大させていったのか。
c. どのような理論言語学的要因で新形式はその使用を拡大させていったのか。
d. 新形式を用いる選好度合いはイ形容詞ごとに異なるのか。

本稿は、このような先行研究において取り残されている問いに対し、予備的な考察を

* 本研究は2020-2021 年度研究活動スタート支援「敬語表現の選択:コーパスを用いた一般化階層ベイズモデリングの理論言語学への統合(代表:山田彬尭)」(#20K21957)の支援、および2021-2026年、研究拠点形成事業(先端拠点形成型)「自然言語の構造と獲得メカニズムの理解に向けた研究拠点形成(代表:宮本陽一)」(#JPJSCCA20210001)の助成を受けた研究成果の一部である。

与える。状態空間モデルという統計手法を、歴史コーパスから得られたサンプルに対して適用し、得られた定量的な統計解析の結果をもとに、以下の点を主張する。

(3)　本稿の主張

a. 新形式が大きくその使用を拡大させるのは 1930 年代前後であるが、その存在がコーパス上に散見され始めるのは 19 世紀末である。

b. その表現が生起するジャンルにより構文の選好度合いが変わる。国定教科書では旧形式への指向が高く、文芸では新形式への指向が高くなる。

c. 認識的モーダル「ょう」や終助詞が共起することで新形式への指向が強まる。

d. 「有難い」「めでたい」という一部のイ形容詞では旧形式への指向が高い。これは、旧形式が新形式では表せないイディオム的意味を持つためではないか、と考えられる。その他の形容詞については、本稿で調査を行ったコーパスからは、明瞭な結果は得られなかった。

2. 先行研究

「国家語」の建設を目的とし規範的文法の制定が政策によって進められた明治時代に、大規模な方言調査に基づいた規範文法書である『口語法』(國語調査委員會(編) 1916)が編集される。その記述から 20 世紀初頭の規範意識が窺い知れるものと期待されるが、果たして、この『口語法』には(1)b に見た「ございます」へ接続するイ形容詞が紹介されている一方で、「です」に直接連なる(1)a の用法については掲載が認められない。このことから、現在(21 世紀初頭)の規範意識とは異なり、当時は、新形式に規範性が認められていなかったことが推察される(川口 2014)。

川口(2014:78-88)は、この(1)b の旧形式を規範的だとする記述は、20 世紀前半に共通して見られる特徴であり、正式に新形式が規範として認められるのは 1952 年の第 1 期国語審議会の「これからの敬語」における記述を待たなければならないことを指摘している。この国の決定を受け、規範文法書においても、翌年に出版された三上章の『現代語法序説―シンタクスの試み―』(三上 1953)や、三尾砂の『話しことばの文法』(三尾 1958)などから徐々に規範形としての位置づけが記載され始めるようになる。また、国定教科書についても同様の傾向が認められる。1904 年から 1949 年まで六期にわたって刊行された国定国語教科書の中で新規形式が記載されているのは、第六期(昭和 22 年、1949 年)の記述だけであり(「いいです」という例が数例採取される)、ここから、1950 年代をこの形式が規範として確立される時期として認めることが妥当であると、川口は結論

付けている（川口 2014:88-92）。

もちろん、規範文法書の執筆者の目に留まるようになったのは、それ以前にすでにこの新形式の使用頻度が増加をしてきていたからこそであり、1950 年代に先立って「美しいです」型の丁寧語構文が人々の間で用いられ始めていたことが推察される。先行研究では「慣用化」が進んだ時期と、「規範化」した時期を分けたうえで、1950 年代を新形式が「規範化」した時期、1930 年代を「慣用化」した時期だと位置づけることが妥当だという見解が提案されているが（川口 2014:88）、しかし、このように「慣用化」や「規範化」が進行した時期については大まかな成立年代についての予測がついている一方で、それでは具体的にいつ頃からこの表現が登場し始めたのかという「黎明期」に関する研究・検証はこれまでのところ十分になされているとは言い難い。

そこで本稿では、この「慣用化」期と目される 1930 年代を含む明治から昭和期をカバーする日本語歴史コーパス（CHJ）を対象に、そこに登場している(1)に掲げられた両形式の使用の移り変わりを、時系列分析の一種である状態空間モデル（State-Space Model）を用いて解析し、いつ頃からこの表現が登場し、また、どのような要因が存在するときに、そして、どのようなイ形容詞が用いられた時に、新／旧形式の選択割合が高まるのかを、定量的な調査から浮かび上がらせ、その結果を記述・報告する。

3. 状態空間モデル

状態空間モデルは、観測されるデータの背後に、観測はされない（できない）ものの観測データを生み出す確率を支配する潜在的な状態 (State)が存在し、それが時系列に沿って変化をしていくという仮定を持つ統計モデルである（Shumway and Stoffer 2017; Durbin and Koopman 2001; Hagiwara 2021, a.m.o）。本稿が対象とするような通時的構文交替においては、二つの構文の生起が観測データ（従属変数）として設定される。旧形式（「美しゅうございます」型）を 0、新形式（「美しいです」型）を 1 として表現すれば、（共時的）構文交替はベルヌーイ分布に従う確率的挙動として理解される。すなわち、採取された i 番目の文の丁寧語形式y_iは、パラメータがπの同一のベルヌーイ分布から独立にサンプリングされた標本であるとみなすことができる。状態空間モデルでは、このような観測標本とパラメータとの間を表す式を「観測モデル」と呼ぶ。

(4) $y_i \sim Bern(\pi)$ (Model 1 out of 3)

実際には、このπの値は、様々な独立変数$x_1, x_2, \cdots, x_p$の値によって変化するであろう。

また、分析結果に影響を与えるデータ間の相関に対応するために、グループ因子を変量効果として投入することも必要である。例えば、旧形式をとるのか新形式を取るのかという選択傾向は、用いられたイ形容詞ごとに変わる可能性がある。これらの点を踏まえ、ロジスティック回帰分析の枠組みに沿い、対応する P 個の偏回帰係数、切片、変量効果をそれぞれ$\beta_p(p \in \mathbb{N}_1^P)$、$\beta_0$、$u_{0j}$ $(j \in \mathbb{N}_1^J)$と表現し(J は考察にのぼるイ形容詞の数)、π を次のように線形に分解して表現をする(F^{-1}には通例何らかの累積分布関数が用いられるが、ここでは最も無標な想定であるロジスティック関数の累積分布関数を仮定する)。

(5)
$$y_{ij} \sim Bern(\pi_{ij}) \qquad \text{(Model 2 out of 3)}$$
$$\pi_{ij} = F^{-1}(\eta_{ij})$$
$$\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_P x_{Pij} + u_{0j}$$
$$u_{0j} \sim N(0, \tau^2)$$

さらに、どちらの構文が選択されやすいかという選択確率は、年に応じて変化していくことが想定される。この想定をこの統計モデルに反映させる場合、もはや単一のβ_0を考えるだけでは不十分で時点 t におけるβ_0を考えることが必要となる。これを$\beta_0{}^{(t)}$とおく。この$\beta_0{}^{(t)}$と過去の時点の状態$\beta_0{}^{(t-1)}, \beta_0{}^{(t-2)}, \beta_0{}^{(t-3)}, \ldots$との関係を記述したものが「状態モデル」であり、本稿では、次の式に示される、切片に一次のマルコフ性を持たせるシンプルな状態空間モデルを分析に仮定し、その結果を解釈していく。

(6)
$$y_{ij}^{(t)} \sim Bern\left(\pi_{ij}^{(t)}\right) \qquad \text{(Model 3 out of 3)}$$
$$\pi_{ij}^{(t)} = F^{-1}\left(\eta_{ij}^{(t)}\right)$$
$$\eta_{ij}^{(t)} = \beta_0^{(t)} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_P x_{Pij} + u_{0j}$$
$$u_{0j} \sim N(0, \tau^2)$$
$$\beta_0^{(t)} \sim N\left(\beta_0^{(t-1)}, \sigma^2\right)$$

4. データ

4.1 コーパス

ここでは、国立国語研究所によって構築された「日本語歴史コーパス(CHJ)(2021.3版)」の中で明治期から昭和期に属する事例を分析する。中納言(2.5.2 版)から次の検索式を用い対象となる言語表現を抽出した上で、変量効果への推定精度を高めるため、5 回以上登場したイ形容詞 1,031 例(32 のイ形容詞)に対象を限定し分析を進める。

(7)　旧形式(「美しゅうございます」型)

キー: 品詞 LIKE "形容詞%" AND 後方共起: 語彙素="御座る" ON 1 WORDS FROM キーAND 後方共起: 語彙素="ます" ON 2 WORDS FROM キー

(8)　新形式(「美しいです」型)

a. キー: 品詞 LIKE "形容詞%" AND 後方共起: 語彙素="です" ON 1 WORDS FROM キー

b. キー: 品詞 LIKE "形容詞%" AND 後方共起: 語彙素="た" ON 1 WORDS FROM キー AND 後方共起: 語彙素="です" ON 2 WORDS FROM キー

4.2 変数とパラメータの推定

日本語歴史コーパスで提供されているジャンル情報、形態素情報を利用し、表 1 に示された変数を(6)で示された統計モデルに投入し、このモデルに出現するパラメータを推定する。パラメータの推定は、ハミルトニアン・モンテカルロ法によって事後分布を近似するアルゴリズムが搭載された Stan (Stan Development Team 2020)を用いて行われた(操作は R から行う; R Core Team 2020)。それぞれの固定効果を含めるか否かでモデルには2^5個の可能性が存在するが、本研究ではこれらのモデルをすべて推定した後、「一つ抜き交差検証法情報量基準(Leave-One-Out cross-validation Information Criterion, LOOIC)」と「広く使える情報量基準(Widely Applicable Information Criterion, WAIC)」を用いモデル比較を実施した(Watanabe 2010; Vehtari et al. 2016, 2017)。結果、最もパフォーマンスがよいモデルとして判断された「過去時制x_5以外の全ての変数を投入したモデル」を最良のモデルとして選択し、以下の議論ではこのモデルの推定結果に基づいて解釈を行う(ここでは紙幅の都合上 WAIC の結果のみを図 1 に表示しているが、LOOIC も非常に類似した結果となっている)。

なお、事前分布には Stan のデフォルトの指定を踏襲し、MCMC の反復回数を 25,000 回(うちバーンイン期間は 18,000 回)とした上で、MCMC サンプルの自己相関を減らすため間引き(thinning)の大きさを 15 として推定を行った。チェーンはデフォルト(4 つ)の設定で実装し、各パラメータに対して得られる(25,000-18,000)/15 * 4 = 1,868 個のサンプルを基に事後分布に関する各種統計量を計算する。パラメータの有効サンプルサイズの最小値は 1010.414 であり、50%を超える値が得られた今回のケースではサンプルの効率性についても問題はないものと判断をした(1010.414/1868 * 100 = 54.1%)。また、収束の指標となる$\hat{R}$の値も全て 1.01 以下の値を取ったことを確認した(Vehtari et al. 2021)。

変数	種類	内容
構文の選択	従属	その観測事例が旧形式なら 0、新形式なら 1 の値を取
y	変数	る二値の離散的な従属変数。
教科書	固定	その観測事例が「国語教科書」ジャンルに属するなら 1
x_1	効果	を、それ以外では 0 を返すダミー変数。
文芸	固定	その観測事例が「文芸」ジャンルに属するなら 1 を、そ
x_2	効果	れ以外では 0 を返すダミー変数。（なお、参照ジャンル
		は「非文芸」である。）
終助詞	固定	その観測事例が終助詞を伴っているなら 1 を、それ以
x_3	効果	外では 0 を返すダミー変数。
認識的モーダル	固定	その観測事例が認識的モーダル「ょう」を伴っているな
x_4	効果	ら 1 を、それ以外では 0 を返すダミー変数。
過去時制	固定	その観測事例が過去時制（例：「美しかったです」）な
x_5	効果	ら 1 を、それ以外では 0 を返すダミー変数。
変量効果	変量	固定効果では捉えられない j 番目の形容詞の独自性
u_{0j}	効果	を表す変数。$N(0,\tau^2)$に従うと想定する。

表 1 モデルに投入された変数.

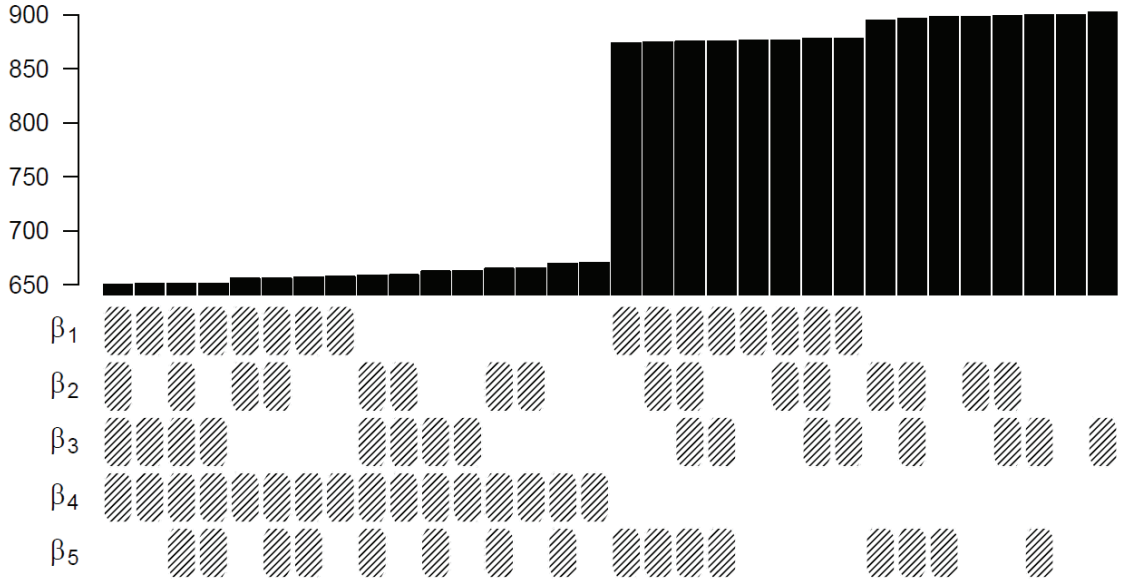

図 1 WAIC に基づくモデル比較. 下部のバーコードプロットはどの偏回帰係数がモデルに組み込まれているのかを表し、上部の棒グラフはそのモデルに対して計算された WAIC の値を表示している。

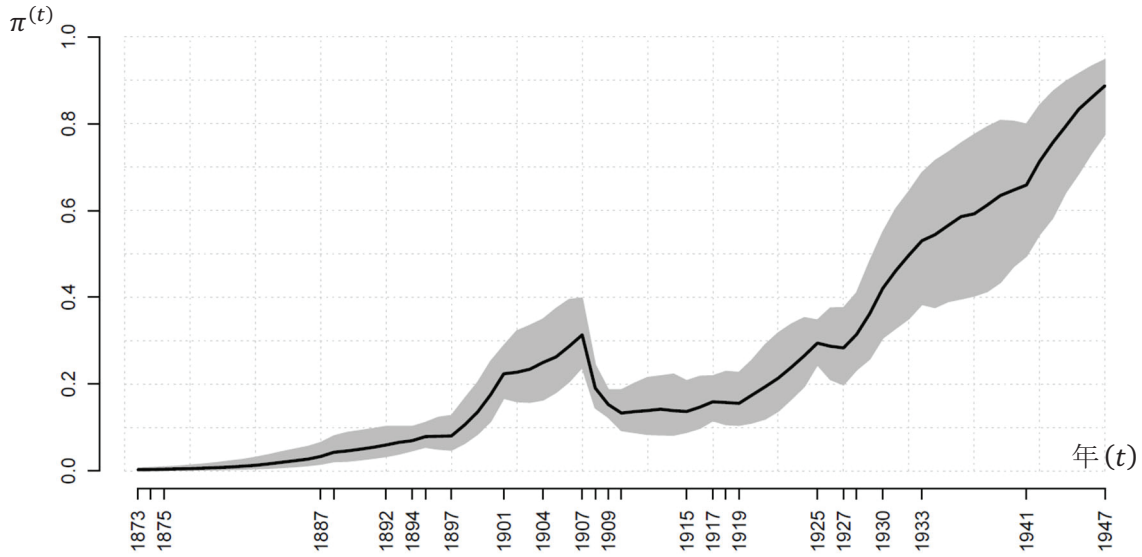

図 2 各年における$\pi^{(t)}$の事後分布．実線は事後中央値をつないだもの、灰色の領域は 99%ベイズ確信区間を表している。

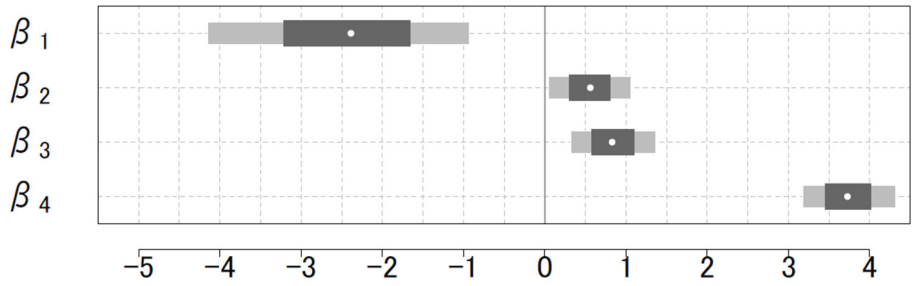

図 3 固定効果の偏回帰係数の事後分布．白丸は事後中央値、濃い帯は 66%ベイズ確信区間、薄い帯は 99%ベイズ確信区間を示している。

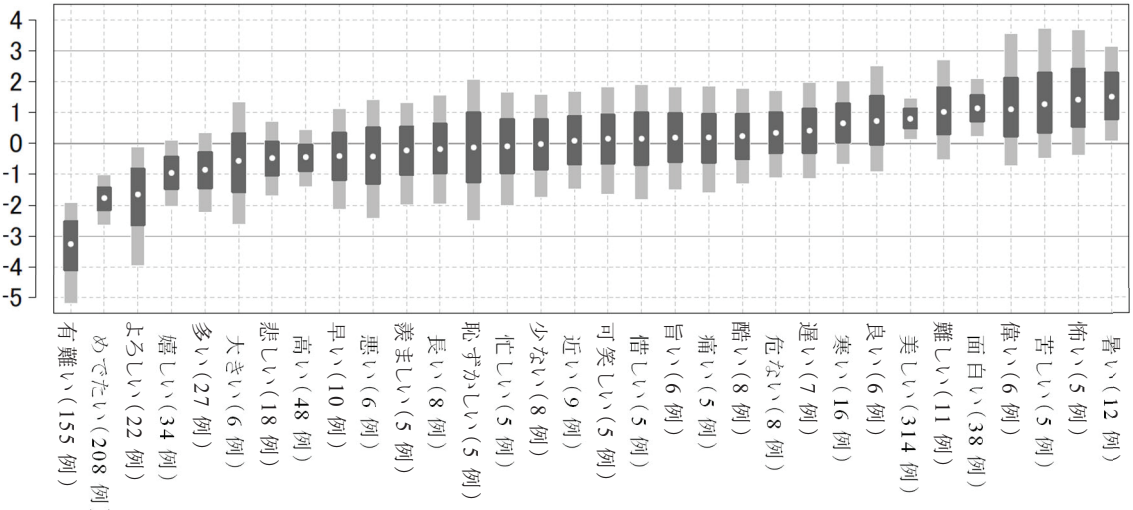

図 4 変量効果u_{0j}の事後分布．各形容詞の独自性を表す変量効果を事後中央値（白丸）の値の小さい順から並べたもの。濃い、薄い帯は 66%、99%ベイズ確信区間。

5. 結果と解釈

5.1 通時的構文選択の変遷

図 2 はそれぞれの年に対しての$\pi^{(t)}$の事後分布を時系列に沿って提示したものである。川口が「慣用化」期と捉えた 1930 年代あたりから大きく「美しいです」型の構文選択の比率が大きく上昇していることが読み取れ、先行研究の指摘を定量的（視覚的）に裏付けられる結果を得た。しかし、新形式の使用は、さらに時代をさかのぼった 19 世紀末から徐々に見られていたことが併せて明示的に示されており、およそ 50 年ほどの時間的幅を持ちつつ次第に新形式の定着化が進んでいった過程が読み取れる。

5.2 固定効果の分析1：社会言語学的要因（ジャンルの影響）

四つの固定効果に対する推測結果は、図 3 に示されている。最も大きな負の効果量を持つ変数として「教科書」の存在が認められ、一方「文芸」では正に振れた位置に事後分布が形成されていることから、参照ジャンルとなった「非文芸」ジャンルに比べ、教科書では旧形式が、文芸ジャンルでは新形式が指向されていることが分かる。国定教科書は規範性が高く旧形式への指向が高いことは予想されることではあるが、教科書ではない他のジャンルにおいて、教科書の記載傾向よりもはるかに多く新規形式が使われていたことが、この結果から客観的に裏付けられたことになる。

5.3 固定効果の分析2：理論言語学的要因

次に、統計モデリングの結果、理論言語学的要因として考察対象に据えられた要因の推定結果について考察する。

第一に、最も大きな正の効果量を持つ変数は、認識的モーダル「ょう」であり、ここから「美しゅうございましょう」よりもはるかに「美しいでしょう」という表現が選択されやすかったことが窺い知れる。

第二に、終助詞の存在も正の効果量を示しており、「美しゅうございます／美しいです」と「美しゅうございますね（よ/…）／美しいですね（よ/…）」を比べると後者の方に、新規表現を指向する効果があったことが認められる。

第三に、図 1 に見るように「過去形」の有無は、ベストなモデルやセカンドベストのモデルに含まれているわけではなく、この変数にモデルのパフォーマンスを高めるような効果がないことが示唆される。すなわち、「美しゅうございます／美しいです」と「美しゅうございました／美しかったです」の違いは、旧形式と新形式の選択にあまり関わっていないということが窺い知れる。

5.4 変量効果の分析：形容詞間のバリエーション

変量効果を含めたモデルを作る利点の一つが、イ形容詞間での新形式を取る選択傾向の差を分析できることであり、この事後分布の推定結果を示したものが図 4 である。

まず、「有難い」と「めでたい」は事後分布が大きく負に振れており、旧形式への依存度が他の形容詞に比べて高いことが分かる。これは、「ありがとうございます／（お）めでとうございます」に「ありがたいです／めでたいです」では表しきれないイディオム的な意味が存在し、そのため旧形式での使用相対頻度が高かったためなのではないか、と考えられる（なお「早い」については 10 例しか用例がないこと、そして、その中に占める非イディオム的意味での割合が大きいことがあり、顕著な旧形式指向性は見られなかった）。

次に、新規形式を指向した形容詞について考えると、本データからは明瞭な傾向は観察されなかった。確かに、事後中央値に注目すると「暑い」「怖い」などが他の形容詞より大きな値を示しているが、確信区間が広いため、確定的な判断を下すことは控えるべきであろう（この確信区間の大きさは同コーパスにおける当該構文での使用が少ないことに起因する）。先行研究では、第六期国定教科書の「イ形容詞＋です」形が全て「いいです」に限られていた（川口 2014）という指摘があり、この形容詞が強い新形式指向性を見せるかと予測されたが、これも確信区間の幅が広く今回の事例からは確定的な診断が難しい。99%確信区間が 0 を含まず、事後分布が正に振れている事例には「面白い」「美しい」が存在するものの、他の形容詞の確信区間を踏まえると、これら以上に大きな値を取る形容詞が存在する可能性も現時点では否めない。将来の研究では、より精度の高い推定を実施するために、さらに大規模なコーパスを用いる必要があろう。

6. まとめと今後の課題

先行研究においては「美しいです」型の新形式の「慣習化」が 1930 年代に、「規範化」が 1950 年代に進んだという見解が示されているが（川口 2014）、具体的にいつ頃から、どのように、この表現が使われ始めたのかという「黎明期」に関する調査は進んでいなかった。この課題を解決するため、本研究では、通時コーパスの定量分析を通じ、(i) 新形式の使用がコーパス上に散見され始めるのが 19 世紀末からであること、(ii) 国定教科書では旧形式への指向が高く、文芸では新形式への指向が高いこと、(iii) 認識的モーダル「ょう」や終助詞が共起することで新形式への指向が強まること、(iv) 「有難い」「めでたい」についてはイディオム的意味を持つことから旧形式への依存が高いこと、を明らかにした。なお、紙幅の都合上、これらの記述的な傾向の分析が言語変化や丁寧語の言語理論に対して、どのような理論的な示唆を持つのかについては十分に論じるこ

とができなかった。これについては稿を改めて論じることとしたい。

参考文献

Durbin, James and Siem Jan Koopman (2001) *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Hagiwara, Junichiro (2021) *Time Series Analysis for the State-Space Model with R/Stan*. Singapore: Springer.

川口良 (2014)『丁寧体否定形のバリエーションに関する研究』くろしお出版.

國語調査委員會（編） (1916)『口語法』 國定教科書共同販賣所. DOI: 10.11501/1870063

三上章 (1953)『現代語法序説―シンタクスの試み―』刀江書院.

三尾砂 (1958)『話しことばの文法』法政大学出版会.

R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. Https://www.R-project.org/.

Stan Development Team (2020) Stan Modeling Language Users Guide and Reference Manual. version. 2.21.0, https://mc-stan.org.

Shumway, Robert H. and David S. Stoffer (2017) *Time Series Analysis and its Applications: with R examples*. New York: Springer.

Vehtari, Aki, Andrew Gelman and Jonah Gabry (2016) *loo: Efficient leave one-out cross-validation and WAIC for Bayesian models. R package version 0.1.6.* https://github.com/stan-dev/loo.

Vehtari, Aki, Andrew Gelman and Jonah Gabry (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27. 1413–1432.

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter and Paul-Christian Bürkner (2021) Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis* 16(2). 667–718.

Watanabe, Sumio (2010) Asymptotic equivalence of Bayes cross validation and Widely Applicable Information Criterion in singular learning theory. *Journal of Machine Learning Research* 11. 3571–3594.