日本語文法研究はデータをどのように扱ってきたか

中俣(なかまた)　尚己(なおき)　（大阪大学）

1.　はじめに

このシンポジウムの趣旨・概要の冒頭は「過去の学会誌や研究会誌を眺めていると、現在とは異なる分析手法や分析内容があることに気づかされる。」で始まっている。また、発表者は中俣(2024)において、2022 年-2023 年における日本語文法（理論・現代）の展望論文を書いたが、そこでも理論的道具立てや調査方法に工夫がなされた研究が多く勉強になった。発表者は、たまたまこれまでに調査が行われていない現象を調査することは進化とみなさず、新しい方法論のおかげで、これまでとは異なる分析が可能になったり、新しい現象に光があたるようになることを進化とみなしたい。

発表者は理論言語学には明るくなく、自身ではコーパスを用いた量的研究に取り組んでいる。そこで、発表者は『日本語文法』掲載の研究論文を対象に、過去の研究がどのように数値と向き合ってきたかを調査し、そこから進化の流れを論じようと考えた。

2.　方法

2001 年の創刊号から 2025 年の第２号までの「研究論文」計 314 編を対象にした。寄稿論文、研究ノート、書評は対象外とした。

調査は「どのように数値を扱っているか」である。文法研究には大別して「作例」に基づくものと「実例」に基づくものがある。しかし、これは車の両輪であり、作例が実例になったから進化したとは言えない。一方、作例のみに基づいた研究であっても、容認度調査を元にモデル化を行い、どのような要因が解釈に強く影響しているのかを特定するような研究(井上ほか 2022)などは新しい方法論を切り開いていると評価できる。また、コーパスを使ったとしても、単に例文をあげているだけなら、それはその例文を見つけるのにかかる時間が短縮されたということなので、あくまでも数値を扱っているかどうかを調査した。

表 1 のように整理すると全体の中で(c)と(d)がどれぐらいあるかということを調査したということである。

表１　数値データによる研究の分類

	実例に基づく研究	作例に基づく研究
数値データを扱わない研究	(a)実例をとりあげる	(b)発表者の文法性判断のみで論じる
数値データを扱う研究	**(c)条件に合致する例の数を報告する**	**(d)文法性判断に関する調査結果を報告する**

※数値データを扱う研究には、心理実験やテキスト全体の計量研究も一部含まれる。

3.　結果

3．1　全体による傾向と経年変化

まずは、各年の調査結果を図 1 に示す。

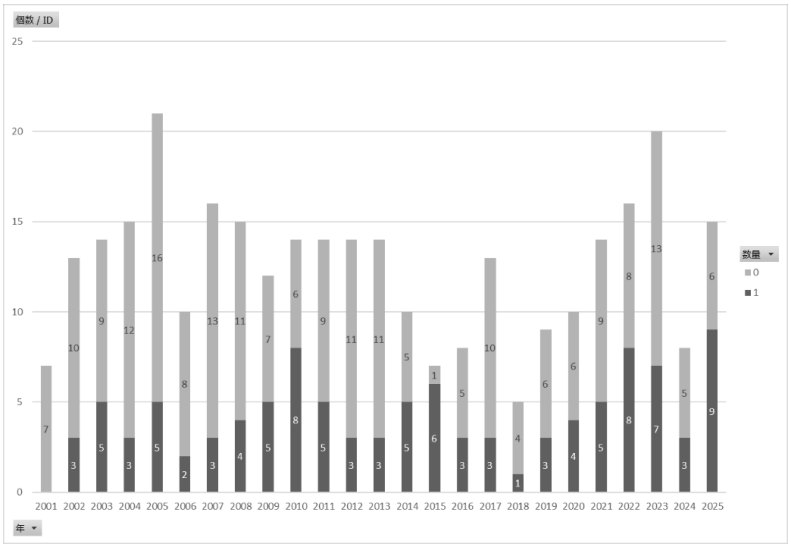

図 1　25 年間の全論文のうち、数値を扱ったもの(=1)

図 1 からすぐに読み取れることは『日本語文法』においては数値を扱った研究の割合はほとんど変化がないということである。コーパス自体は 2011 年の BCCWJ 公開前後から利用が増えている。しかし、例の数を報告するような研究の増加には結びついていない。ただし、統計的には 2013 年以前よりも 2014 年以降の方が数値を扱った研究が多い。数値を扱った研究が 50%を超えたのは 2010 年、2015 年、2025 年の 3 回のみである。数値データを扱った論文は全部で 106 本であり 314 例中 34%に相当する。

なお、2016 年の日本語文法学会では「文法性判断」がテーマとして取り上げられた。背景にはコーパス研究の波の中で、文法性判断による研究の重要性を再確認するという意味があったと思われる。しかし、掲載論文を見れば、このタイプの研究の割合は減少していない。

3．2　作例研究における数値の報告の少なさ

次に、この数値を扱った研究は実例か作例かというと、そのほとんどは表 1 の(c)に該当する実例の例文数の報告である。文法性判断について、例えば X 人中 Y 人がその文を OK と判断したという記述があれば、それは数値を報告したとみなす。そのような基準でカウントした、表 1 の(d)に該当する論文の掲載数を図 2 に示す。25 年間で文法性判断の数値を報告した研究は 11 本のみであった。

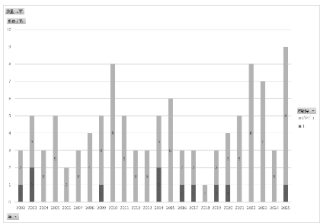

図2　数値を扱った研究における文法性判断の割合(黒)

3．3　数値の処理の仕方

数値の報告といっても、ただ文を正しいと判断した人数や用例数(raw data)を報告するだけでなく、その割合を報告するなどの統計的操作を行ったものも存在する。数値を扱った 106 本の論文のうち、どのような統計的操作を行ったかで集計したものが表２である。

なお、ここで記述統計について補足しておく。一般には記述統計とは平均や標準偏差など、多数のデータの特徴を圧縮して示すための処理であり、要約統計量と呼ばれることもある。しかし、例文数などのカウントデータにおいては平均値の算出は馴染まない。割合の計算は P=Y/X という形であり、X と Y という２つのデータを圧縮・要約しているとみなせるため、ここでは記述統計に含めた。実際には記述統計にカウントしているのはほとんどが割合の計算である。「文法研究は全く統計を利用していない」という結論を避けるため、このような処理を行った。

表2　数値を扱った論文における統計処理の内実（重複あり）

大分類	処理の目的	具体例	論文数(N=106)
記述統計	多数の数値を要約し、データの特徴をわかりやすくする。	割合、平均値、相関、標準偏差、特化係数	61(58%)
推測統計	少数のデータから、全体の傾向を確かめる。特に群間に差があるのかを検証する。	t 検定、χ^2 検定、分散分析	13(12%)
多変量解析	様々な大量のデータから、傾向や特徴を発見する。	主成分分析、対応分析、クラスター分析	2(2%)
モデリング	ある変数の値が決まった時に結果を予測するモデルを作る。	ロジスティック回帰、決定木	1(1%)

割合の計算、つまり実数に加えてパーセントの値まで出しているような研究は６割に満たない。発表者の感覚からはこれは少ないと感じる。パーセントの値がないと、35 例中７例と 59 例中 10 例ではどちらのほうが多いか、読者としては直感的にわかりにくい。このような時にパーセント表示があればと思う。だが、４割ほどの研究がそれを行っていない。また、時代による変化も見られない。昔からきちんと割合を出している研究は出しているし、現在でも出していない研究は出していないということである。

また、多変量解析とモデリングは発表者も近年取り組んでいる手法であり、文法研究にも新たな光を当てるものと期待している。（推測統計は心理学や計量言語学分野ではすでに下火である。）しかし、『日本語文法』においてはこの手法を扱った論文は３本のみである。また、その３本中２本の著者に玉岡賀津雄氏が入っていることには触れておくべきだろう。（玉岡 2005,　李ほか 2023）

3．4　分野による傾向差

以下、あくまでも参考程度であるが発表者の主観によって研究を「現代・理論・歴史・方言」の４つに分類し、分野ごとに数値を扱った研究の割合を示したものである。歴史的研究は基本的に用例数を報告するものが多い。ただし、記述統計（割合）を示さない研究は依然として見られる。また、現代語については、傾向は全論文の傾向と一致した。

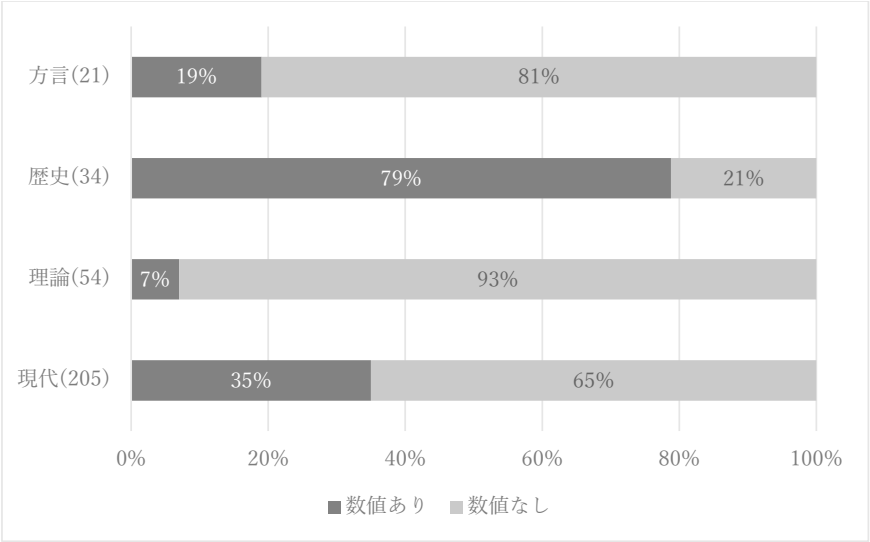

図３　研究分野ごとの数値を扱った研究の割合

4．　今後の日本語文法研究

4．1　実例に基づく研究

コーパス研究は基本的に例の数を数えることにより研究を進めていく。現在でも、例を取り上げるだけの論文もあるものの、歴史的研究を中心にすでにこの方法は定着しているといえる。しかし、実数（＝粗頻度）を報告するにとどまる研究が依然として多いことは気にかかる。数の大小を論じるためには必ず比較という心的操作を必要とする。そのためには単位をそろえる必要があり、割合を計算してパーセントで示すといったことが必要になる。このような心的操作を読者に任せてはいけない。

割合の計算、パーセントの示し方は小学校で学習する内容ではあるが、卒論指導をしていても適切な表・グラフを作れない学生は非常に多い。論文の査読においてさえ、初歩的なミスは散見される。中俣(2021)はパーセントの計算だけを扱った非常に初歩的な言語統計学の入門であるが、学生指導には役立つものと考える。パーセントは分母を 100 とした値であるが、必ずしも 100 でなくてもよい。史的研究の分野では分母を 10 とした研究も存在する(森 2023)。

また、異なるコーパスないしサブコーパスを比較するとき必ず調整頻度を算出する必要があることを忘れてはならない。この時の単位にはよく 100 万語あたりの語数(pmw)が使われるが、数値は何でもよい。もちろん、コーパスの総語数というのはテキストの選定、語の認定といった大きな問題が絡むが、しかし、誤差はあれど傾向差を明らかにすることは重要である。例えば、「中納言」の「まとめて検索 KOTONOHA」という機能は検索した語がどのコーパスに多いかを図示してくれる非常に便利な機能を有しているが、これが可能になるのも pmw という共通の単位を用いているためである。ある単一のコーパスについて、形式 X が N 例出現したという情報は単独ではなんの価値ももたない。しかし、pmw という共通の単位を多くの研究者が用いれば、異なる論文どうし、異なる資料どうしであっても、数値の比較が可能になり、相違または変化を論じられるようになってくる。2010 年代は多くの文法研究者が BCCWJ を使った時代であった。単位は統一されていなくても、資料が統一されていた。しかし、今後はより多様なコーパスを複合的に観察することが必要になってくる。その際には単位をそろえることを常に念頭におかなければならない。

4．2　作例に基づく研究

発表者が危うさを覚えるのは、25 年間で文法性判断における数値を報告した論文が 11 例しかない、という事実のほうである。研究者ひとりの内省のみで良いということは、他の研究者や学生がどんな判断を持ってきても、「その例文、他の人にも聞いてみた？」「もっと多くの人に聞いてみたら？」と言うことはできないということになる。この言葉を言ったことがない人間のみが、自身の内省のみに基づいて論文を書く資格がある。

もし、そうではなく、多くの協力者に対して文法性判断を行ってもらっているというのであれば、その事実は数値とともに報告するべきである。『日本語文法』の論文を読むと「査読者から例文の解釈について疑問が付されたが、執筆者の語感では問題ない」と押し切ったり、「執筆者の周囲に聞くとほとんどが同じ判断であった」と曖昧に書くケースも見られる。しかしながら、5 名中 4 名のほとんどなのか、50 名中 40 名のほとんどなのかは、統計学的には全く異なる意味をもつため、数値の報告は必要であると考える。

発表者は作例に基づく研究は文法研究において欠かすことができないと考えている。コーパスには限界があるからだ。発表者が問題視しているのは（よりにもよって利害対象者本人の）ひとりの内省だけで論を進めることである。文法性にせよ、意味解釈にせよ、自分以外の話者も自分と同じ判断であるという担保が必要であるという主張である。心理学者が自身の内省のみに基づいて心理効果を主張したという話は聞いたことがない。また、研究で使用するすべての例についてアンケートを行えという主張でもない。しかし、研究においてキーとなる例文は必ずあるはずである。その例文についてだけでも、調査を行うべきであるという主張である。

調査を行う際には条件をきちんと統制し、ある程度まとまった数の協力者に対して調査を行うということである。この点で、2016年のシンポジウムで取り上げられた上山・傍士(2017)、Hoji(2015)の調査システムは卓越している。この研究ではある種の文法性判断ができる話者とできない話者がいるという前提に立ち、できる話者を選別するようにデザインされている。この文法性判断ができる話者とできない話者という考え方については金水(2017)は「ステレオグラム」のたとえを出して説明している。

発表者はまず多くの協力者に質問をするという点に強く賛同する。一方で、文法性判断はできる、できない、訓練によって身に着ける、という考え方よりも、そこには個人差があり、異なる文法が共存していると考えるほうが「記述的」であると考える。言語が変化することを考えれば、すべての時空間で文法性判断が一致することはあり得ない。共時的に見れば、調査を行えば必ずどこかに不一致が見られると考えるべきであろう。

このような状況においても、綿密に用意された例文を多数の協力者に回答してもらい、理論に合わない結果を切り捨てるのではなく、すべての回答データをまるごと使って統計モデリングを行うことで、どのような要因が文法性判断に影響しているのか、また個人差はどれぐらいあり、回答者にはどのようなパターンが存在するのかを明らかにすることが可能になる。

図4はモデリングでよく使われるロジスティック曲線である。この曲線自体は言語変化のS字カーブ(Aitchison1991)として説明に用いられることが多いが、基本的には0から1の値を出力する関数のグラフである。この性質を利用して、ある文が容認される確率を表現することも可能である。発表者は2016年のシンポジウムで傍士氏に「文法性判断は3段階ぐらいまでならできると思うが、4段階、5段階というのは可能だろうか？」という質問を投げかけた。氏の答えは「訓練を積めばできる」ということであった。しかしながら、訓練をつまなくても、多数のデータを元にロジスティック回帰分析を行うことで、ある文の文法性（あるいは容認される確率）を0から1の間で予測したり、記述したりすることも可能になる。つまり、文法性判断があいまいな文というのは、100%正文でも100%非文でもないということであるが、このパーセントは何かの量ではなく、確率と考えるのだ。実際には条件が整えば言えるが、整わなければ言えないといったことであろうが、これは確率の考え方にマッチする。

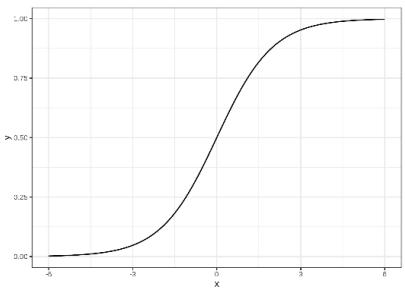

図4　ロジスティック曲線

実際にロジスティック回帰分析を応用した研究としては井上ほか(2022)を挙げることができる。量化子の解釈について1文につき13もの質問を重ねた研究であるが、どのような因子がどの程度効いているのかを一般化線形混合モデル(GLMM)によって明らかにしている。

4．3　オープンサイエンス

少し方向性は変わるが、オープンサイエンスという概念についても触れたい。研究成果を専門家・非専門家問わず誰でもアクセスできるようにする、という試みである。オープンサイエンスの構成要素のうち、論文を無料で公開するオープンアクセスについては、我々にとってもかなり身近な存在になってきたと言ってよいであろう。ただし、日本語文法学会ではこの問題を議論するのは難しい。

ここでは、得られたデータを公開するオープンデータや、方法論を公開・共有するオープンメソドロジーについて研究者が意識すべきと思われる。この点については方言研究者の営みが大きな参考になる。『岡崎敬語調査』や『方言文法全国地図』の回答が Excel ファイルで公開されており、統計的な分析を行える準備が整っている。また、松岡葵氏の『日琉諸語の調査票ポータルサイト』などはオープンメソドロジーの優れた取り組みである。

発表者の取り組みはそれと比べれば小規模であるが、多変量解析の分析結果を付録として公開するようにしている。実際、中俣(2020)の付録データを元に、馬場(2021)が別角度からの検証を行っている。また、分析用のコードも公開している。研究者のためというわけではいが、授業でコーパスを分析した結果をハンドブックの記事として公開する『文法コロケーションハンドブック E』プロジェクトも、研究成果を全世界の学習者に伝えるという観点からはオープンサイエンスと呼べるかもしれない。

発表者は現在、院生たちと Baayen(2008)を使って統計と R の勉強会を実施している。この本には LanguageR という R のライブラリが含まれており、その中には、反応時間や親密度調査の回答結果も含んだ様々な研究の生データが含まれている。読者はそのデータを使って統計計算やグラフ描画を行っていくのだが、例えばオランダ語の接辞や語源に関するデータ、ソロモン諸島の言語のデータなど、学生にとって、とっつきにくいデータも多い。

文法研究者が綿密に計画された調査を行い、そのデータや調査票を公開することで、妥当性の検証や研究者育成につながると考えられる。文法性判断の個別の回答など、何の役に立つかわからないデータであっても、例えば 25 年もすれば、言語変化に関連する貴重なデータになる可能性はある。この時、生の回答データであれば、時間にかかわる変数を一列加えるだけで、すぐにモデリングを行って時間の影響を測定することが可能になるのである。

5．　おわりに

本発表では過去 25 年間の『日本語文法』掲載論文を調査し、数値の扱いについてはほぼ何の変化も見られないということを明らかにした。しかしながら、文法性判断に対して統計モデリングを適用することで、文法研究を進化させることができると考えられる。さらに、データの積極的な公開は、研究者コミュニティ全体の進化につながることを主張した。

参考文献

井上雅勝・藏藤健雄・松井理直(2022)「日本語量化文の解釈と処理方略」『言語研究』162, 91-118

上山あゆみ・傍士元(2017)「容認可能性と言語理論の説明対象」『日本語文法』17-2, pp.20-36.

金水敏(2017)「文法研究におけるデータについて―文法研究は経験科学たりうるか―」『日本語文法』17-2, pp.54-63.

玉岡賀津雄(2005)「中国語を母語とする日本語学習者による正順・かきまぜ語順の能動文と可能文の理解」『日本語文法』5-2, pp.92-109.

中俣尚己(2020)「主成分分析を用いた副詞の文体分析」『計量国語学』32-7, 419-435.

中俣尚己(2021)「言語統計学入門(3) ―パーセンテージと比率―」『計量国語学』33-3, pp.205-213

中俣尚己(2024)「2022 年・2023 年における日本語学界の展望　文法（理論・現代)」『日本語の研究』20-2, pp.21-28.

馬場俊臣(2021)「副詞の文体の計量について―― 中俣論文の主成分分析結果との比較 ――」『語学文学』60, pp.26-35, 北海道教育大学語学文学会.

森勇太(2023)「江戸後期洒落本に見る丁寧語の運用とその地域差―京都・大坂・尾張・江戸の対照」『日本語文法』23-1, pp.104-120.

李依格・張佩霞・玉岡賀津雄(2023)「話し言葉における動詞の否定ていねい形「～ません」「～ないです」を選択する言語外的要因」『日本語文法』23-2, pp.87-102.

Aitchison, J. (1991) *Language change: progress or decay? 2nd ed.,* Cambridge University Press.

Baayen, R. H.(2008) *Analyzing Linguistic Data*, Cambridge University Press.

Hoji, H. (2015) *Language Faculty Science,* Cambridge University Press.

ウェブサイト

国立国語研究所「データとプログラム」
https://www2.ninjal.ac.jp/hogen/dp/dp_index.html

松岡葵「日琉諸語の調査票ポータルサイト」
https://sites.google.com/view/japonicIanguages-questionnaire/

中俣尚己「文法コロケーションハンドブック E」
https://grammarcollocation.wordpress.com/