形態素解析の大規模言語調査データへの応用 ： 岡崎敬語調査パネルデータにおける名詞・代名詞・動詞の相対頻度数に対する話者性別効果の検証

著者	松田 謙次郎
雑誌名	国立国語研究所論集
号	7
ページ	151-165
発行年	2014-05
URL	http://doi.org/10.15084/00000529

国立国語研究所論集 (NINJAL Research Papers) 7: 151–165 (2014)
ISSN: 2186-134X print/2186-1358 online

形態素解析の大規模言語調査データへの応用
――岡崎敬語調査パネルデータにおける名詞・代名詞・動詞の相対頻度数に対する話者性別効果の検証――

松田謙次郎

神戸松蔭女子学院大学／国立国語研究所 共同研究員

要旨

Seifart et al.（2010）および Seifart（2011）は名詞・代名詞・動詞の談話中における相対頻度数（NTVR）が言語内で，また言語間でも大きな分散を示し，類型論的に興味深い分布を示すものであることを明らかにした。ここでは岡崎敬語調査（国語研 1957, 1983, 阿部（編）2010, 西尾他（編）2010, 杉戸 2010a, 2010b, 松田他 2012, Matsuda 2012, 松田他 2013, 井上・金・松田 2013）の回答文に形態素解析を施したデータを分析することで，(1) NTVR が回答者の加齢に影響を受けずほぼ一定の値を保っており類型論的指標として信頼しうる安定性のある数値であること；(2) NTVR には性差が見られ男性の値の方が女性の値より高いこと；(3) この性差が敬語補助動詞の使用頻度の性差によるものであると考えられること，の 3 点を主張する。NTVR は生涯変動を見せない安定した指標であるが，NTVR 算出を目的とした談話データの使用に際しては，当該言語の社会言語学的変異にも配慮する必要がある。また，この研究は形態素情報付き岡崎敬語調査発話データの有用性の一端を示すものであり，こうしたデータの活用によって，岡崎敬語調査のデータは計画当初考えられていたものよりも遙かに多くの多種多様な言語学的問題に解答を与えることが期待される *。

キーワード：岡崎敬語調査，形態素解析，パネルデータ，コーパス，性差

1. はじめに

岡崎敬語調査については，2008 年に第 3 回目の調査が終了して以来多種多様な分析と報告が行われてきており，半世紀を経た岡崎市における敬語変化の全貌が徐々に明らかにされつつある（国語研 1957, 1983, 阿部（編）2010, 西尾他（編）2010, 杉戸 2010a, 2010b, 松田他 2012, Matsuda 2012, 松田他 2013, Inoue 2013, 井上・金・松田 2013, 井上 2013）。岡崎敬語調査はその長い調査スパンもさることながら，話者の発話が回答文として記録してある点がその大きな特徴の一つである。すなわち言語行動調査においては，特定状況下において話者がどのような発話をするかを問い，その回答をそのまま回答文として記録しているわけである。過去の岡崎敬語調査の分析はこの回答文における敬語要素の分析を主としているが，近年急速に発展した日本語形態素解析の技

* 本研究は，2013 年 8 月 12–13 日にドイツ・ライプチッヒ市のマックスプランク研究所で開催されたワークショップ “The relative frequencies of nouns, pronouns, and verbs in discourse. An international workshop” で口頭発表されたものに大幅に手を加えたものである。MeCab と UniDic について御教示くださった田中牧郎先生，原稿にコメントを下さった井上史雄先生，そしてマックスプランク研究所での発表時にコメントを下さった方々に感謝申し上げる。本研究は，国立国語研究所基幹型共同研究プロジェクト「日本語の大規模経年調査に関する総合的研究」（プロジェクトリーダー：井上史雄，2012 年度～）の一部としてなされた研究である。また，2013 年度日本学術振興会科学研究費補助金（B）「変異理論の新展開と日本語変異データの多角的分析」（研究課題番号：25284082，研究代表者：松田謙次郎）の助成を受けている。

術を用いて諸々の情報をタグづけすることで，回答文データを多角的な視点から解析することが可能になる。岡崎敬語調査は元来敬語の変遷の調査を目的としたものであったが，こうして加工されたデータは，半世紀にわたる個人の言語使用の変遷を品詞別から活用形に至るまでの詳細な文法情報を使って解析できるものとなり，その利用価値は飛躍的に高まることが期待される。

本稿では，形態素解析を施された回答文を用いて，現在類型論的見地から注目されている名詞・代名詞・動詞の相対頻度数（noun-to-verb ratio，以後 NTVR）に対する話者の加齢や性別による影響を分析し，NTVR に関する先行研究では顧みられなかった性差が存在することを立証し，それが回答者の性別による敬語補助動詞の使用差によるものであることを主張する。このことは，NTVR 算出のベースになる言語データの採取や分析に際して話者の性別（そしてさらにはその他の社会言語学的諸属性）を考慮する必要があることを意味するものであり，世界中の多種多様な言語を対象とする NTVR 研究に重要な示唆をもたらすものである。また，この作業を通して岡崎敬語調査のデータに形態素解析を施すことで開かれる広大な可能性の一端を示すことにする。

2. 名詞・代名詞・動詞の相対頻度数研究とは

Seifart et al.（2010）および Seifart（2011）は（1）で表されるような名詞・代名詞・動詞の談話中における相対頻度数が言語内で，また言語間でも大きな分散を示し，類型論的に興味深い分布を示すものであることを明らかにした。

（1） 名詞・代名詞・動詞の相対頻度数（NTVR）＝（名詞の数＋代名詞の数）/ 動詞の数 [1]

この研究は，Seifart らによるフォルクスワーゲン財団の危機言語ドキュメンテーションプログラム（Volkswagen Foundation DoBeS program）に端を発するもので，この過程で製作された 7 言語それぞれ 3 万語規模のコーパス解析から NTVR の問題が浮上し，マックスプランク研究所における Seifart とその同僚らによる新たなプロジェクトとして派生したという経緯を辿っている。当初地理的・文化的・類型論的に多様な 7 言語（Baure 語，Bora 語，Chintang 語，Lamunkhin Even 語，N|uu 語，Popopluca 語，Sakha 語）の分析から始まった研究は，調査対象言語を拡張しつつ，その後多種多様な言語において精力的に分析が進められている。とりわけ NTVR 差を産み出すメカニズムをめぐっては，各言語の形態・統語論的特質，当該言語共同体のソーシャルネットワークのサイズ，談話のジャンル差やストラテジー，さらに一般的な人間の認知的制約などさまざまな観点からの可能性が検討されており，同プロジェクトは計算機言語学・コーパス言語学的手法を用いた新たな大型類型論研究として注目を浴びつつある [2]。

[1] なお，研究によっては（名詞の数＋代名詞の数）/（名詞の数＋代名詞の数＋動詞の数）という数式を用いているものもあるが，本論では以下特に断らない限り（1）に示した数式で議論を進める。
[2] プロジェクトの詳細については，フォルクスワーゲン財団へのプロジェクト計画書（Seifart et al., n.d.）および以下の同プロジェクトのウェブサイトを参照のこと：“The relative frequencies of nouns, pronouns, and verbs cross-linguistically” http://www.eva.mpg.de/linguistics/research/typological-surveys/the-relative-frequencies-of-nouns-pronouns-and-verbs-cross-linguistically.html（2014 年 1 月 8 日参照）

たとえばすぐに NTVR に関わりそうな要因として，当該言語において項の実現が義務的であるか否か（± pro drop）という特質が浮かぶが，実はこれが談話中の名詞・代名詞数の予測にはあまり有効ではないという結果が得られている（Seifart 2011）。統語的特性ではむしろ語順の効果が確認されており，動詞後置型言語では非後置型言語に比べて名詞の割合が高いことが判明している（Bickel et al. 2013）。談話的側面に目を向けると，語り（narrative）においては，はじめは名詞の割合が高いものの談話の進行とともに徐々に減少し，その後再び上昇するというサインカーブ的分布パターンが存在する（Seifart 2011）。また，談話の話速（speech rate）も重要な要因であり，話速の遅い部分では名詞の使用が集中するという報告がなされている（Seifart et al. 2013）。さらに社会学的見地からは，話者と聴者の社会的距離が近い場合には両者による共有情報の多さから，名詞の使用頻度の低下が見られ，NTVR の分布が単に言語内的な要因のみでは説明が付かないことが明らかになりつつある（Bickel 2006, 2011）。こうした結果からは，NTVR が言語の複数レベルにわたる複合的作用により現出した現象であることが見て取れよう。

3. 問題点と仮説

Seifart らの始めた NTVR 研究プログラムは，かように言語の複数レベルにわたる総合的課題に，コーパス言語学や計算機言語学的手法を援用することで豊饒な結果をもたらしつつあるが，ここで一つの疑問が浮きあがる。それは，これらの研究の中で，NTVR があたかも各言語で固定した数値として捉えられている点である。言語の話者を集団として捉えた場合，この値はある一定の平均値に収束するとしても，個人レベルでは当然散らばりが予測される。そうした散らばりは個人内の変動，また性差をはじめとする各種話者属性による差異が考えられる。

個人内変動では，個人の成長や加齢に伴う変化がありえよう。言語習得において音韻や文法の基本部分が思春期付近でほぼ完成を迎えることには疑いの余地はないが，文法の固定がそのまま NTVR の固定につながるとは限らない。たとえば Seifart（2011）は談話のジャンルや談話内の位置によってこの数値が変動することを指摘しているが，これはつまりこの数値には談話的要因が考えられるということを意味している。談話レベルでの習得は文法の習得よりも時間を要するはずであり，たとえば若者と老人では談話技術に相当な差が見られるはずであり，それが NTVR の差につながることも予想される[3]。また，日本語であれば加齢に従って敬語使用も変化するはずであるが，それに伴う動詞の使用頻度の変動も容易に予測できる。つまり，成長や加齢によって NTVR が変動すると予測するに十分足るだけの根拠があるわけである。

もちろん変動は個人内ばかりではなく，性差をはじめとする個人の属性差による差異も考えられる。性差は多くの言語変異研究で重要な要因として働いているが，相対頻度数でも話者の性別による差が見られても不思議はない。

さて，こうした変異を検証するためには，同一男女の集団の発話を長年にわたり追跡した大量

[3] たとえば井上・金・松田（2013），Inoue（2013）では「ていただく」に成人後習得が見られることが報告されている。

のデータ，それも発話が文字化された上に各単語の形態素情報がタグづけされたデータがあればよい。残念ながら管見の及ぶ限り既存コーパスではこれらの条件を満たすものは存在しないが，国語研が 1953 年から 3 次にわたり実施してきた岡崎敬語調査のうち，パネルサンプルの発話データ（回答文）に形態素解析を施せばこの条件にマッチするデータを作成することができる。各品詞の相対頻度数を見るためには，形態素情報をタグづけされたデータからそれぞれの品詞を抽出して集計すればよい。パネルサンプルは最長で 55 年にわたる同一個人の発話を追跡しており，これを分析することで上述の問にも十分な答えを出すことができるはずである。

検証作業を進めるにあたって仮説を設定しよう。上にも触れた通り，過去の言語習得の知見から思春期以後の文法はほぼ固定されていると見なしうる根拠があるので，加齢の影響については「NTVR は話者の加齢にかかわらず一定である」という帰無仮説が考えられる。性別については，先行研究で同様な問題を扱ったものがなく，確たる根拠がないことからやはり帰無仮説として「NTVR には，話者の性別による有意な差はない」という仮説を立てるのが望ましい。またこれら 2 つの仮説の組み合わせとして，「NTVR には，性別にかかわらず加齢による差は生じない」という 3 つめの仮説が立てられることになり，結局（2）のような仮説群が設定される。

（2） 岡崎敬語調査パネルデータにおける NTVR の分布に関する仮説
- （a）NTVR は加齢によって影響されることなく一定である。
- （b）男女間において NTVR の差はない。
- （c）NTVR には，性別にかかわらず加齢による差は生じない。

4. データと分析手法：岡崎敬語調査パネルデータ

4.1 データ

データである岡崎敬語調査パネルデータについて説明しよう。岡崎敬語調査は第 1 次〜 3 次の調査ごとに取られた 3 つのランダムサンプル（「継続サンプル」）と，1・2 次継続サンプルの回答者を 19 年後，36 年後に再調査をした 3 つのパネルサンプルから成る（図 1）。これらのデータセットは調査間隔で整理すると，19 年間隔，36 年間隔，55 年間隔の 3 つのタイプに分けられる。このうち 36 年間隔にはもともと第 1 次調査のランダムサンプルから派生したもの（以下「1 次派生」と呼ぶ）と，2 次のランダムサンプルから派生したもの（同様に「2 次派生」とする）があることに注意されたい[4]。無回答が多かったデータを除き，実際に分析に使用可能であったデータの一覧を表 1 に掲げておく。実際の分析に当たっては，たとえば 19 年間隔の変化を比較するのであれば 1 次派生 2 次サンプルの 179 人と，彼らの 1 次継続データでの回答を比較している。他の間隔についても同様である。

[4] なお，岡崎敬語調査の各サンプルについて公式に発表されたサンプルサイズと，ここで記したサイズには若干のずれがある。

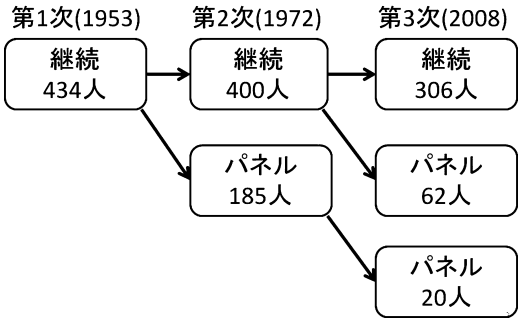

図 1　岡崎敬語調査サンプル関係図

表 1　パネルサンプル・分析回答者数

サンプル	男性回答者数	女性回答者数	合計
1 次派生 2 次	90	89	179
1 次派生 3 次	10	8	18
2 次派生 3 次	23	35	58

4.2 方法論

まずパネルデータから，全話者について生年，年齢，回答文を抽出した。岡崎敬語調査の回答文原データは，すべての回答がカタカナ表記で入力されている。正確な形態素解析のためには漢字仮名交じりの方が望ましいので，一人のアルバイターによる手作業で漢字仮名交じりに変換したものを形態素解析の入力データとして使用した [5]。

形態素解析では MeCab（Ver. 0.995）（http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html）と UniDic（Ver. 2.1.2）（伝 他 2007）の組み合わせで，茶まめ（Ver. 2.0）（小木曽他 2007）を用い，すべての単語について形態素情報を付した。この形態素情報がタグづけされたファイルから名詞・代名詞・動詞を抽出し，仮説に応じた集計を行い，上記の数式で計算した数値を使い t- 検定および分散分析により有意差を検討した。

5. 分析

5.1 加齢効果の分析

まず加齢効果を検証する。図 2 は，19 年，36 年，55 年間隔での NTVR の差をグラフに表したものである。1（つまり名詞と代名詞の和が動詞の数に等しい）を中心に数値は微かに上下し，それぞれのペアで 2 回目の調査ではわずかに NTVR が低い方へと移動しているようである。しかし全体的に見るとこうした変化は微々たるものと言うべきであろう。実際各間隔における数値

[5] 現在鑓水兼貴氏（国語研）によって，岡崎敬語調査回答文すべての漢字仮名交じりバージョンが作成されつつあり，今後公開される見通しである。今回作成したバージョンでは同一語に対する表記が一定していないという欠点があるが，鑓水氏版ではこうした欠点が克服される見通しである。今後の分析には鑓水氏版を使用するのが望ましいことは言うまでもない。

の差を対応のある t-検定で検討したところ，いずれも 5% 域で有意差[6]は検出されなかった（表 2）。つまり NTVR は 19 年，36 年，55 年とその長さにかかわらずほぼ一定の値を保っており，きわめて安定した指標であると言える。

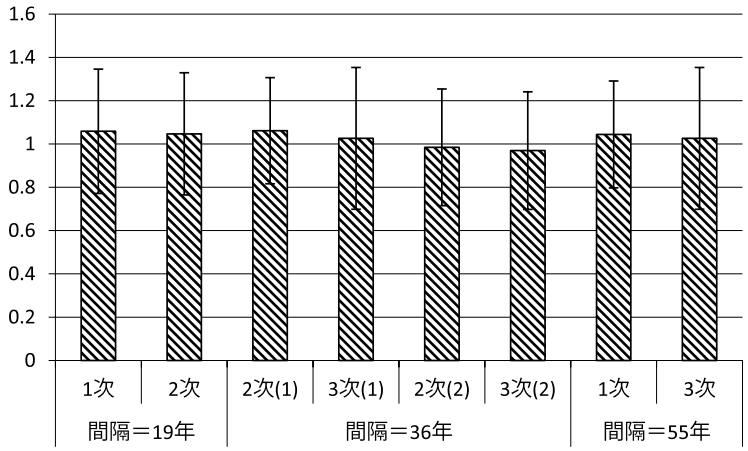

図 2　各間隔における NTVR の変化[7]（※エラーバーは標準偏差を示す）

表 2　各間隔における NTVR の変化の検定結果

間隔	ペア	t 値	自由度	$p <$（両側）
19 年	1 次 /2 次	.421	178	.674
36 年	1 次派生 2 次 /3 次	–.229	18	.821
	2 次派生 2 次 /3 次	–.477	57	.635
55 年	1 次 /3 次	.223	17	.826

ところで，一般的に加齢に伴い発話量が増加することが広く知られている。今回のデータで形態素の数を計算すると，やはり各間隔で発話量は増加しており，それぞれについて対応のある t-検定によって有意差が検出された（表 3, 図 3）。すなわち確かに加齢に伴い発話量は増加したが，NTVR はほぼ不変であったことになる。

[6] 本論文では危険率 5% 以下を統計的に有意と判定した。
[7] グラフ中のカッコ内の数字は派生した回次を表す。たとえば「3 次 (1)」は「1 次派生の 3 次パネルサンプル」を表す。

表 3　各間隔における形態素数の変化の検定結果

間隔	ペア	t 値	自由度	$p <$（両側）
19 年	1 次 /2 次	–4.547	178	.001
36 年	1 次派生 2 次 /3 次	–2.916	18	.001
	2 次派生 2 次 /3 次	–6.536	57	.001
55 年	1 次 /3 次	–3.656	17	.001

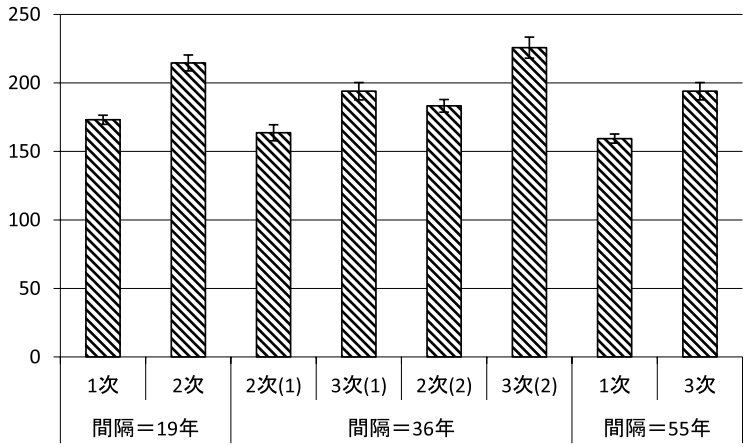

図 3　各間隔における形態素数の変化（※エラーバーは標準偏差を示す）

ここまでの分析は同一回次調査の話者をすべて一括して分析している。しかし回答者の集団は広い年齢層を含んでいる。そこでこのうちもっとも若い話者だけに分析対象を絞れば，言語習得期を終えてから日も浅いので変化の余地があり，結果的に大きな NTVR 差が生じるのではないかという反論が考えられる。この可能性を検証するために，19 年，36 年，55 年の各間隔で，間隔の始まりの調査時（たとえば 1 次 /2 次であれば 1 次の段階）に 30 歳未満であった回答者を抽出し分析を行ったが，対応のある t- 検定で有意差は見出されなかった（表 4)。つまり，若い層だけを取り出しても，NTVR に加齢効果は認められないわけであり，やはり NTVR は言語形成期以降，ほぼ一定であると結論づけることができる。

表 4　各間隔若年層（30 歳未満）回答者における NTVR の変化の検定結果

間隔	ペア	t 値	自由度	$p <$（両側）
19 年	1 次 /2 次	–.821	55	.415
36 年	1 次派生 2 次 /3 次	.329	11	.748
	2 次派生 2 次 /3 次	–1.213	18	.241
55 年	1 次 /3 次	.322	11	.754

5.2 性差

次に仮説（b）の性差効果を検討しよう。図 4 は 19 年，36 年，55 年間隔のデータについて，性別の分布をグラフにしたものである。

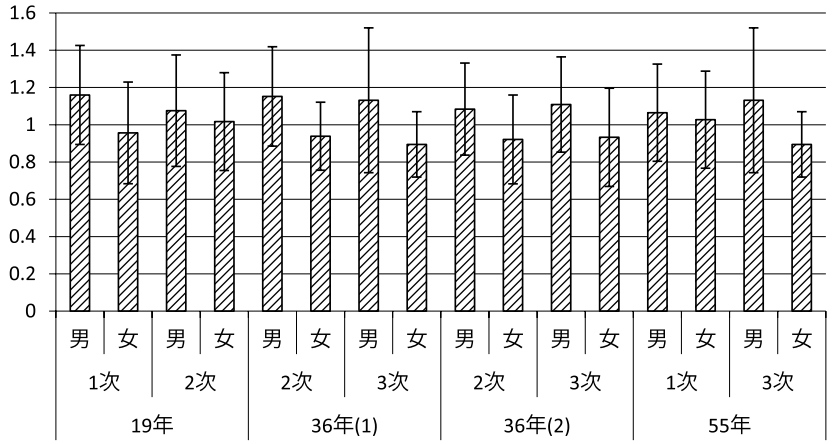

図 4　各間隔における性別による NTVR の変化（※エラーバーは標準偏差を示す）

グラフでは回次ごとに一様に男女差が観察され，差の大小はあれ概ね NTVR は男性＞女性という関係にあることがわかる。検定を施すと，1 次派生 2 次と 1 次派生 3 次を除いた 3 つのデータセットにおいて有意差が認められる（表 5）。

表 5　回次ごとの性別による NTVR 差の検定結果

	t 値	自由度	$p <$（両側）
1 次	5.054	177	.001
1 次派生 2 次	1.399	177	NS
2 次派生 2 次	2.507	56	.016
2 次派生 3 次	2.514	56	.015
1 次派生 3 次 [8]	1.589	18	NS

こうした回次ごとの性差の存在は，先行研究ではまったく触れられていなかったものであり，非常に興味深い。NTVR にこのような一貫した男性＞女性という男女差が生じるメカニズムについては後の議論に譲ることにして，まずここでは NTVR に男女差が存在することを確認しておこう。ここで 2 つ目の仮説であった性差に関する帰無仮説は棄却され，「NTVR には性差による差が存在し，男性の方が女性よりも高い」と結論づけられることになる。

[8] 1 次派生の 3 次データについては，サンプルサイズが小さいことを考慮して Mann-Whitney 検定も行ったが，$U = 26.000, p < .238$ でやはり有意ではない。

5.3 加齢と性差の効果

3 つ目の仮説（c）に移ろう。ここまでの検証から，加齢の効果はないことが確認され，性別については 5 つのグループのうち 3 つで男性＞女性という方向の NTVR 差があることが判明した。それでは，加齢効果が性別によって異なるという可能性はどうであろうか。

図 5，6，7，8 では 4 つの間隔グループそれぞれを男女で分け，19 年，36 年，55 年間での変化をグラフにした。36 年間隔の 2 つのサンプルでは男女はほぼ平行線を描き，性差と加齢の効果が独立であることを示している。一方，19 年サンプルと 55 年サンプルでは，前者が相互に歩み寄るのに対して，後者では相離れるという対称的なパターンとなっている。

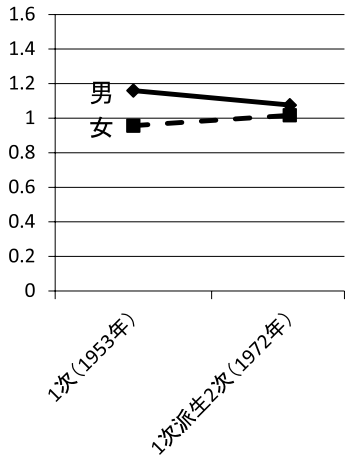

図 5　19 年間隔における性差による NTVR の変化（N：男性＝ 90，女性＝ 89）

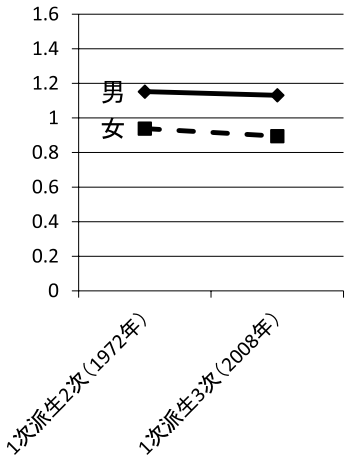

図 6　1 次派生 36 年間隔における性差による NTVR の変化（N：男性＝ 10，女性＝ 8）

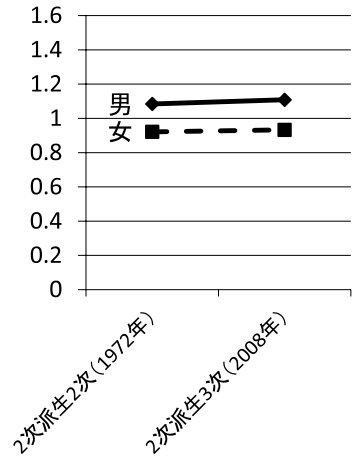

図 7　2 次派生 36 年間隔における性差による NTVR の変化（N：男性＝23，女性＝35）

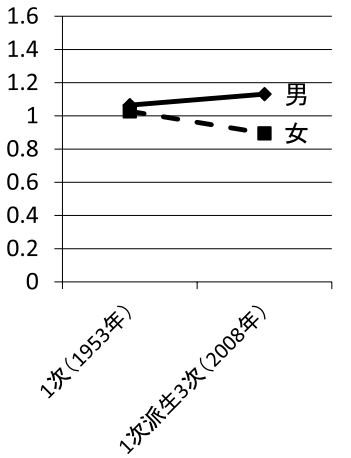

図 8　55 年間隔における性差による NTVR の変化（N：男性＝ 10，女性＝ 8）

NTVR に調査回次（被験者内要因）と性別（被験者間要因）による 2 要因分散分析を施した結果を表 6 に示す。表から明らかな通り，主効果のうち調査回次はすべての間隔で有意でなく，性別は 55 年間隔以外のサンプルすべてで有意であり，19 年間隔では調査回次×性別の交互作用が有意であったわけである。

表 6　3 間隔における回次と性別の分散分析結果

	調査回次	性別	交互作用
19 年間隔	$F(1, 177) = .171$, NS	$F(1, 177) = 19.630, p < .001$	$F(1, 177) = 6.374, p < .02$
1 次派生 36 年間隔	$F(1, 16) = .009$, NS	$F(1, 16) = 6.514, p < .03$	$F(1, 16) = .014$, NS
2 次派生 36 年間隔	$F(1, 56) = .250$, NS	$F(1, 56) = 8.756, p < .005$	$F(1, 56) = .033$, NS
55 年間隔	$F(1, 16) = .110$, NS	$F(1, 16) = 2.147$, NS	$F(1, 16) = .994$, NS

55 年間隔のデータは例外となるが，他 3 つの間隔サンプルでは性差が有意であったことから，加齢効果と同時に考慮しても性差が NTVR の差と関わっていると考えられる。ただし，交互作用が有意であった 19 年間隔サンプルや，交互作用は有意ではなかったが 55 年間隔での男女差のパターンを見ると，加齢と性差の関係はより複雑なものであることも推測できる。

5.1 節での加齢効果の分析と同様，ここでもさらに各サンプル内での年齢層を考慮すべきであるという反論が考えられる。しかし，性別と年齢（ここでは便宜的に 10–20 代の若年，30–40 代の中年，50–60 代の高年と 3 グループに分割するとする）を被験者間要因として信頼性のある分析をするためには，十分なサンプルサイズを持つデータが必要である。そこで今回のデータセット中最大の 179 人のサンプルサイズを持ち，この条件を満たしうる唯一のデータと考えられる 19 年間隔データを使用してこの検証を試みた（図 9）。

グラフでは実線で表した男性のうち高年・中年でともに 19 年間に値が低下する一方，男性若年層と，破線で表した女性のすべての年齢層が同様な動きを呈し NTVR が上昇している。年齢層と性別を被験者間要因，調査回次を被験者内要因とする分散分析では，主効果では性別が $F(1, 173) = 23.147$，$p < .001$ と有意であり，年齢層が $F(2, 173) = 1.898$ と非有意，また被験者内要因の調査回次が $F(1, 173) = .340$, NS と非有意であり，交互作用では調査回次×性別が $F(1, 173) = 9.133$, $p < .004$ のみが有意であった。すなわち少なくとも 19 年間隔サンプルに関する限り，年齢層の違いは NTVR 差に何らの効果をもたらさない一方，性差は有意に NTVR の差を生み，その効果は回次によって異なることが確認されるわけである。

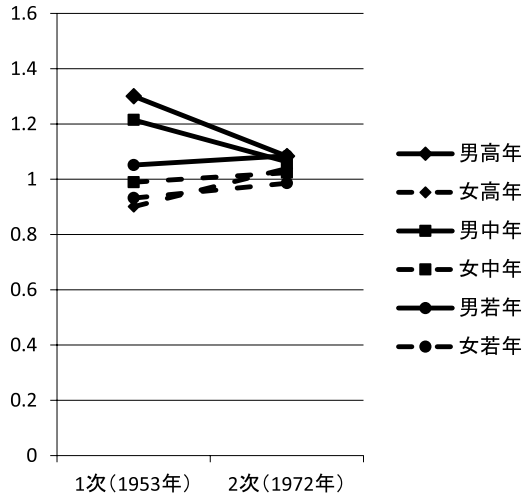

図 9　19 年間隔における性別・被験者年齢層による NTVR の変化（N：男性高年齢層 = 15，男性中年齢層 = 37，男性若年齢層 = 38，女性高年齢層 = 18，女性中年齢層 = 48，女性若年齢層 = 23）

5.4 性差の原因

前節の分析から，加齢と年齢を考慮しても NTVR には性差の効果が認められることが判明した。本節ではその理由を考察してみよう。

男性に比べて女性の NTVR が小さいということは，数式（1）における分子（名詞＋代名詞）の数において男性より女性が少ないか，分母の動詞数が男性より女性が大きいか，又はその両方という可能性が考えられる。実際にデータを検討すると，このうち第 2 の可能性，つまり女性の方が男性より動詞を使用することを示唆するような分布が見られた。これが意味することは，女性の方が男性より多くの文を使用するか，一つの文により多くの動詞を使用するかのいずれかである。文数自体には男女間に大きな違いは見られないため，結局男女差は一つの文の中に使用する動詞の数の違いに求められることになる。文法的に考えると，岡崎敬語調査データで単一文中における動詞数が男女で異なるような事態というのは，敬語補助動詞（「～て下さる」「～て頂く」）の使用頻度がもっとも可能性が高いものである[9]。また，場面の中には「先生」（「この子はあなたのお宅のお子さんです。このお子さんをつれて歩いていると，この人に会いました。この人は，昔あなたが小学校で習った先生です。先生に，「この子は？」とお子さんのことを聞かれたら，何と答えますか。」）のように典型的な回答が「これは私の息子です」のように，動詞を一つも含まないものがあるが，こうした場合でも敬語動詞「ござる」を使用して「これは私の息子でございます」とすることで使用動詞数が増加するというシナリオも考えられる。

[9] 言うまでもなく敬語使用における性差の実証的な報告は，過去の岡崎敬語調査（国語研 1957, 1983）をはじめとして，井出他（1986）など豊富に存在する。岡崎第 3 次調査を終えた段階の報告でも Matsuda（2012），松田他（2012），松田他（2013），Inoue（2013），井上・松田・柳村（2013）などに性差が指摘されている。

この仮説を検証するために各サンプルにおける「て下さる」「て頂く」[10]「ござる」の性別平均使用頻度数を計算した（表 7）[11]。3 語すべてについてすべてのサンプルで男女間に有意差が見られたわけではないが，15 個のセルのうちおよそ半数のセルで有意差が見られたことは注目に値しよう [12]。この他にもテ接続の「てみえる」や,「致す」といった各種敬語補助動詞がデータ中に存在する事実を考えると，男女間の NTVR 差の原因は敬語使用の差に求められる可能性が高いと考えられる [13]。敬語補助動詞をすべて除いて本動詞のみで計算した場合NTVRに男女差は出ないという予想が立つが，この検証は今後の課題としたい。

表 7　各サンプルにおける「て下さる」「て頂く」「ござる」の性別使用頻度数

サンプル	敬語補助動詞	男性	女性	$p <$（両側）
1 次	て下さる	1.411	1.955	.001
	て頂く	.867	1.124	.003
	ござる	.500	1.179	.001
1 次派生 2 次	て下さる	1.272	2.341	.001
	て頂く	.840	1.341	.003
	ござる	.590	.978	.009
2 次派生 2 次	て下さる	1.434	2.457	.009
	て頂く	1.043	1.543	NS
	ござる	.391	.657	NS
1 次派生 3 次	て下さる	2.090	2.500	NS
	て頂く	.909	1.375	NS
	ござる	.818	.375	NS
2 次派生 3 次	て下さる	1.435	2.171	NS
	て頂く	1.391	2.057	.035
	ござる	.522	1.142	.01

[10] 岡崎敬語調査における「て頂く」の増加については，井上（2013）に詳細な報告がある。
[11] 実際には，各サンプルのオリジナルのファイルに含まれているカタカナ版回答文から，「テクダサ」「テイタダ」「ゴザイマ」を抽出した。
[12] ただし 15 回の検定を行っていることから第 1 種の過誤を防ぐために Bonferroni の修正を適用し $\alpha = .05/15 \fallingdotseq .003$ とすると，有意差が見られるのは 1 次の「て下さる」と「ござる」，1 次派生 2 次の「て下さる」の 3 つになる。
[13] 日本語においては一定の語用論的・統語論的条件下において主語名詞句や目的語名詞句などが「省略」され，さらにこうした現象に性差を認める研究（Shibamoto 1985）もあることから，本論で確認された NTVR における性差の原因は，むしろこうした動詞項省略の男女差に求めるべきではないかという反論も考えられる。そもそもこうした「名詞句省略現象」に性差が存在するのかどうかについては Fry（2003）の反論もあるが，NTVR の性差が「名詞句省略現象」によるものであるという可能性は検証するに足るものである。そこで 2 次派生 2 次サンプルについて，回答文で省略されたと判断されたすべての主語名詞句を補充し，新たな（代名詞＋名詞）数に基づいた NTVR を計算し，改めて t- 検定を施した。その結果，性差は $p < .003$ レベルで有意であり，男女間の NTVR 差は「名詞句省略現象」に還元し得ないことがわかる。

6. 結論

以上の考察から, 以下のように結論づけられる。まず, NTVR は加齢によっては変動せず, 19 年, 36 年，55 年間隔のいずれにおいても NTVR はほぼ一定である。これは若年層話者に限定しても変わらず，NTVR は思春期後に固定されると考えられる。また NTVR には性差が見られ，各サンプルにおいてNTVRは男性の方が女性より高い場合が多い。加齢とともに考慮した場合でも，55 年間隔以外では有意な効果が見られたことから，NTVR には何らかの形で性差が関わっていると考えられる。敬語補助動詞の使用頻度の分布から，性差は男女における敬語使用の差異に求められると結論づけられる。

多面的な分析が進められつつある NTVR ではあるが，性差の存在が確認されたのは本研究が初めてであり，さらにその原因が敬語使用の差に求められたことは注目に値する。それはこれが NTVR 値の研究を進める際のデータの問題に関わるからである。本稿の結論からは，日本語の敬語のように，大きな社会言語学的変異を示す文法的システムを持つ言語では，今回同様男女どちらの話者のデータを取るかにより NTVR に無視できないような差が出てしまうわけである。特にこれまでの NTVR 研究が調査してきたような研究蓄積の比較的浅い諸言語では，こうした社会言語学的変異に気付かずにデータ分析を行ってしまい，誤った結論に導かれる可能性も否定できない。ある言語を代表するような談話サンプルをどのようにして採取すべきなのかは簡単な議論で済ませられる問題ではないが，少なくともある言語で NTVR を計測するために談話資料を用いる場合には，その言語において性差をはじめとする社会言語学的変異が確認されるような文法現象の有無が精査されるべきであろう。

7. おわりに

本稿はまた，形態素解析により形態素情報がタグづけされた岡崎敬語調査の発話データの有用性を雄弁に物語るものと言える。最大で 55 年の長きスパンにわたり個人の固定的状況における発話を記録した岡崎敬語調査の発話データは，もちろんそれだけでも高い学術的価値を持つデータであり，その可能性にはまだまだ利用されつくしていないものもある。しかしデータ抽出となると，カタカナ表記のプレーンテキストである現在の発話データでは，せいぜい文字連鎖を対象としたものに限定されてしまう。ここに形態素情報が付与され，それを用いたデータ抽出が可能になると，データとしての有用性は飛躍的に高まることになる。形態素情報のタグづけされた 55 年にわたる発話データなど，管見の及ぶ限り未だ世界のどこにも存在しないデータであり，そこから得られる知見の中には，社会言語学の領域をはるかに越えて心理言語学や社会学にも及ぶものが容易に期待できるはずである。こうした広大な可能性を考えた場合，今回の NTVR 値をめぐる考察はむしろかなり初歩レベルのものと言うべきであろう。

参照文献

阿部貴人(編)(2010)『敬語と敬語意識―愛知県岡崎市における第三次調査―』平成 19(2007)年度～平成 21(2009)年度 文部科学省科学研究費補助金［基盤研究(A)課題番号:19202014, 研究代表者:杉戸清樹］研究成果報告書 第 2 分冊 経年調査基礎データ編.

Bickel, Balthasar (2006) Referential density in typological perspective. Plenary talk, Leipzig Spring School on Linguistic Diversity, March 22, 2006. Leipzig. http://www.uzh.ch/spw/bickel/presentations/rd2006.ppt.pdf.(2014 年 1 月 8 日参照)

Bickel, Balthasar (2011) The role of genealogical units in explaining linguistic distributions: A case study on referential density. Talk given at Workshop "Cross-linguistic and language-internal variation in text and speech: Focus on the joint analysis of multiple characteristics", Freiburg, Germany, February 9–11, 2011. Freiburg. http://www.uzh.ch/spw/bickel/presentations/pears2011.pdf.(2014 年 1 月 8 日参照)

Bickel, Balthasar, Jan Strunk, Frank Seifart, Brigitte Pakendorf, Alena Witzlack-Makarevich, Swintha Danielsen, Søren Wichmann, and Taras Zakharko (2013) Noun-to-verb ratio and grammar. Paper presented at the international workshop "The relative frequencies of nouns, pronouns, and verbs in discourse", Leipzig, August 12–13, 2013.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵(2007)「コーパス日本語学のための言語資源:形態素解析用電子化辞書の開発とその応用」『日本語科学』22: 101–122.

Fry, John (2003) *Ellipsis and wa-marking in Japanese conversation*. New York: Routledge.

井出祥子・荻野綱男・川崎晶子・生田少子(1986)『日本人とアメリカ人の敬語行動―大学生の場合―』東京:南雲堂.

Inoue, Fumio (2013) A contemporary history of Okazaki honorifics—Democratization and te- itadaku—. *Working Papers from NWAV Asia-Pacific 2*, Article 5. http://www.ninjal.ac.jp/socioling/nwavap02/Inoue-NWAVAP2-2013.pdf(2014 年 2 月 5 日参照)

井上史雄(2013)『岡崎「ていただく」の増加』岡崎敬語調査資料集 1(私家版資料集).

井上史雄・金順任・松田謙次郎(2013)「岡崎 100 年間の「ていただく」増加傾向―受恵表現にみる敬語の民主化―」『国立国語研究所論集』4: 1–25.
http://www.ninjal.ac.jp/publication/papers/04 /pdf/NINJAL-Papers0401.pdf(2014 年 1 月 8 日参照)

井上史雄・松田謙次郎・柳村裕(2013)『岡崎敬語「丁寧さ」の変化』岡崎敬語調査資料集 2(私家版資料集).

国立国語研究所(1957)『敬語と敬語意識』東京:秀英出版.

国立国語研究所(1983)『敬語と敬語意識―岡崎における 20 年前との比較』東京:三省堂.

Matsuda, Kenjiro (2012) What happened to the honorifics in a local Japanese dialect in 55 years: A report from the Okazaki Survey on Honorifics. *University of Pennsylvania Working Papers in Linguistics* 18(2), Article 7. http://repository.upenn.edu/pwpl/vol18/iss2/7(2014 年 1 月 8 日参照)

松田謙次郎・阿部貴人・熊谷智子・片岡邦好(2013)「ワークショップ 岡崎敬語調査報告―パネルサンプルの分析―」『日本語学会 2013 年度春季大会予稿集』43–60.

松田謙次郎・阿部貴人・辻加代子・西尾純二(2012)「ワークショップ 岡崎敬語調査報告―継続サンプルの分析―」『日本語学会 2012 年度春季大会予稿集』37–54.

西尾純二・辻加代子・久木田恵(編)(2010)『敬語と敬語意識―愛知県岡崎市における第三次調査―』平成 19(2007)年度～平成 21(2009)年度文部科学省科学研究費補助金［基盤研究(A)課題番号:19202014, 研究代表者:杉戸清樹］研究成果報告書 第 4 分冊 記述調査編.

小木曽智信・小椋秀樹・伝康晴(2007)「日本語研究に適した形態素解析ソフトウェア―「UniDic」と「茶まめ」―」『日本語学会 2007 年度秋季大会予稿集』255–262.

Seifart, Frank (2011) Cross-linguistic variation in the noun-to-verb ratio: The role of verb morphology and narrative strategies. Poster presented at the Association for Linguistic Typology 9th Biennial Conference, The University of Hong Kong, July 21–24, 2011.

Seifart, Frank, Hans-Jörg Bibiko, Balthasar Bickel, Swintha Danielsen, Roland Meyer, Sebastian Nordhoff, Brigitte Pakendorf, Alena Witzlack-Makarevich, Taras Zakharko and NN. (n.d.) "The relative frequencies of nouns, pronouns, and verbs cross-linguistically". Application for Volkswagenstiftung Förderinitiative Documentation of Endangered Languages—Dokumentation bedrohter Sprachen (DoBeS). https://www.eva.mpg.de/lingua/pdf/research/NTVR_Application_Public.pdf(2014 年 1 月 8 日参照)

Seifart, Frank, Roland Meyer, Taras Zakharko, Balthasar Bickel, Swintha Danielsen, Sebastian Nordhoff, and Alena Witzlack-Makarevich (2010) Cross-linguistic variation in the noun-to-verb ratio: Exploring automatic tagging and quantitative corpus analysis. Paper presented at the DoBeS Workshop "Advances in Documentary Linguistics" Nijmegen, October 14–15, 2010.

Seifart, Frank, Jan Strunk, Balthasar Bickel, Brigitte Pakendorf, Alena Witzlack-Makarevich, Swintha Danielsen, Taras Zakharko, and Søren Wichmann (2013) Noun-to-verb ratio and speech rate. Paper presented at the international workshop "The relative frequencies of nouns, pronouns, and verbs in discourse", Leipzig, August 12–13, 2013.

Shibamoto, Janet S. (1985) *Japanese women's language*. Orlando, Fla.: Academic Press.

杉戸清樹 (2010a)『敬語と敬語意識―愛知県岡崎市における第三次調査―』平成 19 (2007) 年度～平成 21 (2009) 年度文部科学省科学研究費補助金［基盤研究 (A) 課題番号：19202014，研究代表者：杉戸清樹］研究成果報告書 第 1 分冊 経年調査実施資料編.

杉戸清樹 (2010b)『敬語と敬語意識―愛知県岡崎市における第三次調査―』平成 19 (2007) 年度～平成 21 (2009) 年度文部科学省科学研究費補助金［基盤研究 (A) 課題番号：19202014，研究代表者：杉戸清樹］研究成果報告書 第 3 分冊 発表成果編.

Analyzing Large-scale POS-tagged Language Survey Data: A Case of Sex Effects on the Noun-to-verb Ratio in the OSH Panel Data

MATSUDA Kenjiro

Kobe Shoin Women's University / Project Collaborator, NINJAL

Abstract

Seifart et al. (2010) and Seifart (2011) calculated the relative frequencies of nouns, pronouns, and verbs (noun-to-verb ratio, or NTVR) in spoken corpora of diverse languages, revealing drastic typological differences. Although the exact reasons for these differences remain unknown, Seifart and his colleagues' innovative line of research has uncovered a number of intriguing grammatical and discourse correlates. Based on statistical analyses of the part-of-speech (POS) tagged versions of panel data from the Okazaki Survey on Honorifics (OSH) (NLRI 1957, 1983, Abe 2010, Nishio et al. 2010, Sugito 2010a, b, Matsuda et al. 2012, Matsuda 2012, Matsuda et al. 2013, Inoue, Kim & Matsuda 2013), we claim that (1) NTVR remains stable for individuals after adolescence, indicating that it is a reliable typological index; (2) NTVR exhibits variation based on speaker sex, with male speakers showing higher values than females; and (3) this sex difference is traceable to a difference in the use of honorific verbs, with female speakers using more auxiliary honorific verbs than male speakers. We conclude that while these results confirm the stability of NTVR within the lifespan of individual speakers, researchers should also take into account the sociolinguistic dimensions of a language when sampling data for NTVR research. Moreover, the analysis demonstrated that the POS-tagged version of the OSH data is a rich source of linguistic information that enables linguists to answer far more diverse questions than the original survey organizers intended.

Key words: Okazaki Survey on Honorifics, morphological analysis, panel data, corpora, sex differentiation