

言語変種を表す複合語「～訛（り）」の諸相 —「NDL Ngram Viewer」と 「国立国会図書館デジタルコレクション」を用いた素描—

岡田 祥平（新潟大学教育学部）

Japanese Compounds Ending in *-namari* Denoting Language Varieties: A Preliminary Survey Using the NDL Ngram Viewer and the National Diet Library Digital Collections

Shohei Okada (Niigata University)

要旨

本発表では、国立国会図書館のデジタル化資料に付与された全文テキストデータを対象に、言語変種を表す複合語「～訛（り）」の諸相を検討する。国語辞典における立項状況を確認したうえで、「NDL Ngram Viewer」と国立国語研究所が構築、公開している各種コーパスを用いて表現を抽出・分類し、さらに「国立国会図書館デジタルコレクション」の検索結果件数から資料上の出現状況の一端を観察した。その結果、国語辞典に立項されていない多様な表現が確認され、この形式が一定の生産性を有する可能性を示した。また、地域変種を表すものが多く、社会変種を表すものは相対的に少ないこと、言語変種そのものだけでなく中間言語的な変種を指す用法もあること、検索結果が一部の分類および表現に集中することを指摘した。なお、本発表は、国立国会図書館のデジタル化資料のうち、全文テキストデータが付与された資料を日本語研究に利用する可能性と留意点を提示するものである。

1. はじめに

日本語では、「お国訛（り）」「東北訛」のような複合語によって、特定の言語変種を表すことがある。これらは、言語変種を指示する語を前項要素とし、「訛」または「訛（り）」を後項要素とする（本発表では、以下、この種の複合語を「～訛（り）」と呼ぶ）。この種の表現は、たとえば、発表者の手近にある小説の中にも、「東北訛」「九州訛」¹、「加賀訛」²、「大阪訛（り）」³、「関西訛（り）」「東北訛（り）」⁴、「関西訛（り）」「津軽訛（り）」「土地訛（り）」⁵など、種々の表現が確認できる。

しかし、そもそも、どのような言語変種が「～訛（り）」と表現されるのか、詳らかではない。この種の問題を取り扱った先行研究・文献も、発表者の気付く範囲では見いだせない。

そこで、本発表では、国立国会図書館のデジタル化資料のうち、全文テキストデータが付与された資料（以下、この種の資料を「NDL 全文テキストデータ」と表記）を対象に、言

¹ 松本清張「皿倉学説」（『別冊文芸春秋』1962年12月号掲載）より。発表者は『松本清張全集38』（文芸春秋社・1974年）に所収されているものを参照した。

² 松本清張「大黒屋」（『オール読物』1964年1月号・2月号掲載）より。発表者は岡本綺堂ほか『時代小説の楽しみ④ 八百八町捕物控』（新潮社・1990年）に所収されているものを参照した。

³ 中川透（＝鮎川哲也）「ペトロフ事件」（『別冊宝石』第8号・1950年）より。

⁴ 鮎川哲也『偽りの墳墓』（文芸春秋新社・1963年）より。

⁵ 内田康夫『死者の木霊』（栄光出版社・1980年）より。発表者は『新装版 死者の木霊』（講談社文庫・2016年）を参照した。

語変種を表す複合語「～訛(り)」の諸相について素描を試みた結果を報告する。まず、2節では、代表的、あるいは日本語使用社会に浸透している言語変種を表す複合語「～訛(り)」を確認すべく、「ジャパンナレッジ Lib」を利用し、『日本国語大辞典 第二版』(小学館)と『デジタル大辞泉』(小学館)で当該複合語の立項状況を確認した結果を報告する。続く3節では、「NDL 全文テキストデータ」を検索するためのプラットフォームの一つである「NDL Ngram Viewer」と、国立国語研究所が構築し、公開している各種コーパスを利用し、現代日本語において、言語変種を表す複合語「～訛(り)」にはどのようなものがあるのかを整理したうえで、2節で報告した二つの国語辞典の立項状況と対比させつつ、小考を試みる。そして、4節では、3節での小考を踏まえつつ、「NDL 全文テキストデータ」を検索するためのプラットフォームの一つである「国立国会図書館デジタルコレクション」を利用し、言語変種を表す複合語「～訛(り)」の資料上の出現状況の素描を試みる。最後に、5節で、まとめを行う。

なお、「NDL 全文テキストデータ」を主な調査対象の資料としたのは、3.2で述べるとおり国立国語研究所で構築された種々のコーパスでは本発表の目的を達成するためには十分な用例を確保できなかったのに対し、後述するように「NDL 全文テキストデータ」では非常に多くの用例を確認できたことによる。また、本発表の主たる目的は言語変種を表す複合語「～訛(り)」の諸相を報告することにあるが、副次的には、岡田(印刷中)でも提案した「NDL 全文テキストデータ」を日本語研究に利用する方法を実践した事例報告としての意味合いも持つ。

2. 『日本国語大辞典 第二版』(小学館)と『デジタル大辞泉』(小学館)に見る、 言語変種を表す複合語「～訛(り)」

言語変種を表す複合語「～訛(り)」のうち、代表的、あるいは日本語使用社会に浸透しているものはいったいどのようなものなのであろうか。そこで、発表者の所属先の附属図書館が契約している「ジャパンナレッジ Lib」を利用し、『日本国語大辞典 第二版』(小学館、以下『日国』と表記)と『デジタル大辞泉』(小学館、以下『大辞泉』と表記)について、言語変種を表す複合語「～訛(り)」の立項状況を確認した。具体的には、「ジャパンナレッジ Lib」の「詳細(個別)検索」画面で、以下のような条件を設定して『日国』と『大辞泉』を対象に検索を行ったうえで、検索結果から言語変種を表す複合語「～訛(り)」を目視で抽出した⁶。なお、以下のような AND 検索を行ったのは、それぞれの辞典での見出しの漢字表記が、「訛り」と送り仮名が存在しているのか、「訛」と送り仮名が存在しないのか、不明だったためである。

- (a) 検索語「訛」／検索範囲を「見出し」／「条件」を「部分一致」 かつ
検索語「なまり」／検索範囲を「見出し」／「条件」を「後方一致」

以上の検索の結果、以下の(1)から(3)のような複合語が二つの辞典に立項されていることが分かった。なお、以下のまとめでは、後項要素「訛(り)」を省略している。

⁶ この過程で、単純語の「訛り」や、言語変種を表さない複合語(「糸訛り」「言葉訛り」「横訛り」など)は本発表の分析対象から除外される。

(1)日本の特定の地方・地域・都市名の言語変種を指し示しているもの： 15 語

①『日国』『大辞泉』双方に立項：^{あすま}東⁷

②『日国』のみに立項： 越後・江戸・葛西・上方・関西・京・西国⁸・薩摩・信州・
中国⁹・東京・東北・坂東・武州

(2)地方・地域は特定しないが、主に地方の（都市部ではない）言語変種を表すもの： 6 語

①『日国』『大辞泉』双方に立項： 田舎・お国・国・土地

②『日国』のみに立項： 地方・里¹⁰

(3)特定の位相の言語変種を表すもの： 2 語

①『日国』『大辞泉』双方に立項： 里¹¹

②『日国』のみに立項：^{さいとう}西島

以上のような二つの国語辞典の立項状況から、言語変種を表す複合語「～訛（り）」はおおむね(1)(2)(3)のような分類ができそうである（(1)(2)は地域的な変種、(3)は社会的な変種を表現している）。しかし、上記の結果には、1 節で紹介した「加賀訛」「津軽訛（り）」「大阪訛（り）」といった語は見いだせない。つまり、二つの国語辞典での立項状況からは、日本語で使用されている言語変種を表す複合語「～訛（り）」の実態を把握することは難しいと思われる。

3. 日本語で使用されている言語変種を表す複合語「～訛（り）」にはどのようなものが存在するのか

3.1 「NDL Ngram Viewer」を使用した調査

日本語で使用されている言語変種を表す複合語「～訛（り）」にはどのようなものが存在するのであろうか。2 節での試みを踏まえつつその点を把握すべく、発表者は国立国会図書館が 2000 年度から進めている作業によってデジタル化された所蔵資料、中でも全文テキストデータが付与されている「NDL 全文テキストデータ」に着目した。

「NDL 全文テキストデータ」を日本語研究に活用しようとする近年の動き、そして「NDL

⁷ たとえば『大辞泉』には「東国地方の言葉のなまり。京言葉に比べて下品なものとされた。」とある。したがって、日本の特定の地方を指していることが分かる。

⁸ 『日国』には「中国、四国、九州地方に特有の訛り。また、その言葉。特に、九州地方の訛りをさしている場合が多い。」とある。したがって、日本の特定の地方を指していることが分かる。

⁹ 『日国』には「中国地方の言葉に特有の発音。」とある。したがって、ここでいう「中国」とは中華人民共和国のことではないと分かる。

¹⁰ 『日国』では「里訛り」の語釈の一つ目として「さとことば（里言葉）(1)」に同じ。」と記載している。

『日国』の「里言葉」では(1)として「いなかことば。地方なまりのあることばづかい。くにことば。さとなまり。」という語釈を掲載しているのであるが、『大辞泉』にはこの語釈は存在しない。

¹¹ 『日国』では「里訛り」の語釈の二つ目として「さとことば（里言葉）(2)」に同じ。」と記載している。

『日国』の「里言葉」では(2)として「遊里で主として遊女が用いた特別なことばづかい。遊里語。くるわことば。さとなまり。」という語釈を掲載している。一方、『大辞泉』の「里訛り」では「江戸時代、遊里で遊女の使った独特の言葉づかい。郭詞（くるわことば）。里言葉。」という語釈となっている。

全文テキストデータ」の正確性や検索のために利用できる検索プラットフォームについては、岡田（2024，印刷中）にまとめているのでそちらを参照されたい。また、「NDL 全文テキストデータ」を日本語研究に活用する方法については岡田（印刷中）で提案している。ここでは、正規表現を利用して「NDL 全文テキストデータ」を検索できる「NDL Ngram Viewer」を活用して、日本語で使用されている言語変種を表す複合語「～訛（り）」にはどのようなものが存在するかを整理する。

「NDL Ngram Viewer」は NDL ラボ（<https://lab.ndl.go.jp/index.html>）が「NDL 全文テキストデータ」を検索するために開発されたツールである。「NDL 全文テキストデータ」のうち、図書・雑誌資料約 230 万点（図書は主に 1960 年代までに、雑誌は主に 1990 年代までに刊行されたもの）について、検索結果を出版年別に可視化することができるほか、簡単な正規表現を利用した検索も可能である。また、検索結果のデータはダウンロードできる。

本発表では、「NDL Ngram Viewer」を利用し、以下に示す二つの正規表現を用いて検索を行った¹²。

- (b) 「.+訛」： 「訛」の前に 1 文字以上の文字列が存在する
- (c) 「.+訛り」： 「訛り」の前に 1 文字以上の文字列が存在する

その結果、(b)では 5708 文字列、(c)では 1201 文字列を得ることができた。その後、ダウンロードした検索結果のデータを目視で確認し、言語変種を表す複合語「～訛（り）」を抽出し、その前項要素を整理した結果が以下の A から H である。その際、「樟葉訛」「松浦訛」のように、基本的に古典籍にのみ出現するものは除外した。

なお、A から H を参照する際の注意点を四つ、記しておく。第一に、4 節で行う「国立国会図書館デジタルコレクション」での検索との関係上、同一の複合語であっても表記が異なるもの（漢字の異体字の違いは除く¹³）は別の表現として扱っている。つまり、「語彙素」ではなく「書字形」を基準に整理したということである。そのため、以下に示す異なり書字形数を（「語」ではなく）「表現」として数える。第二に、2 節同様、後項要素「訛（り）」を省略して列記している。第三に、四角で囲んだ語は、2 節で確認した二つの国語辞典に立項されていたものである。第四に、以下のまとめは便宜的に行ったものであり、分類の妥当性には検討の余地が残されている。

A 日本の特定の地方・地域・都市などが前項要素となっているもの： 124 表現

蝦夷・北海道・函館・**東北**・みちのく・陸奥・奥州・津軽・青森・羽後・秋田・岩手・盛岡・仙台・庄内・出羽・山形・会津・福島・

¹² (c)の検索は(b)に内包されると思われたのであるが、実際に検索結果を確認すると、(c)の検索結果の中には(b)の検索結果には含まれないものが存在した（「イタリー訛り」など）。このように、「NDL Ngram Viewer」の挙動には外部のユーザーにはうかがい知れない不明な点もあることを記しておく（岡田（印刷中）の「4. 「NDL Ngram Viewer」の活用法」も参照）。

¹³ これは、「NDL Ngram Viewer」でも「国立国会図書館デジタルコレクション」でも、漢字の異体字は検索の際にはその違いが無視されるためである。したがって、たとえば「大阪」には「大阪」という表記はもちろん、「大坂」という表記も含まれる。一方、「浪速」と「浪花」は別表記として扱われる。

関東・**東**¹⁴・**坂東**・常陸・日立・茨城・水戸・栃木・上州・上総・下総・安房・房州・**武州**・埼玉・秩父・**江戸**・**東京**・相模・鎌倉・
 北陸・**越後**・新潟・佐渡・越中・富山・加賀・能登・金沢・越前・若狭・福井・
 甲斐・甲州・信濃・**信州**・佐久・伊那・駿河・伊豆・遠州・三河・尾張・名古屋・木曾・
 美濃・飛騨・伊賀・伊勢・志摩・桑名・
関西・**上方**・近江・江州・京都・**京**・丹後・丹波・但馬・播州・神戸・淡路・難波・浪
 花・河内・大阪・大和・紀伊・紀州・
 吉備・備前・岡山・広島・長州・山口・山陰・出雲・四国・阿波・讃岐・伊予・松山・
 土佐・高知・
 九州・筑紫・筑後・福岡・博多・肥前・佐賀・唐津・長崎・島原・肥後・熊本・天草・
 豊後・大分・日向・**薩摩**・鹿児島・沖縄・
 日本

B 外国の特定の国・地方・地域・都市、あるいは民族などが前項要素となっているもの：

61 表現

アイルランド・スコットランド・イギリス・英国・ロンドン・倫敦・エディンバラ・ヨ
 ークシア・スペイン・フランス・仏蘭西・パリ・巴里・アルザス・ガスコーニュ・ガスコ
 ン・ペリゴール・コルシカ・ドイツ・独逸・ベルリン・ザクセン・シュワーベン・バイエ
 ルン・イタリア・イタリー・伊太利・羅馬・オランダ・スイス・瑞西・ウィーン・ポーラ
 ンド・ウクライナ・スラブ・ユダヤ・ガリラヤ・ロシア・

アメリカ・米国・ニューヨーク・ブルックリン・ボストン・テキサス・ニューイングラ
 ンド・カナダ・キューバ・ラテン・

インド・オーストラリア・朝鮮・京城・支那・蘇州・天津・浙江・湖南・広東・山東・
 山西・

外国

C 方角・方向などが含まれる表現が前項要素となっているもの¹⁵： 11 表現

東国・西・**西国**・西部・中西部・南国・南部・南方・北・北国・北部

D 地方・地域は特定しないが、主に地方部（都市部以外）を意味する語が前項要素となっ
 ているもの： 18 表現

お国・**御国**・**おくに**・**国**・ふるさと・ふる里・古里・故里・故郷・郷土・郷里・**田舎**・
地方・地方的・**土地**・浜・植民地・方言

E 言語名が前項要素となっているもの： 9 表現

英語・スペイン語・フランス語・ドイツ語・中国語・漢語・外国語・日本語・母国語

¹⁴ 「東」は厳密には後述の C に分類すべきかもしれないが、2 節でみたとおり『日国』や『大辞泉』で「東
訛り」の形で「東国地方」で使用されている言語変種を表現する複合語として立項されていたため、あえ
て A として分類した。

¹⁵ ここに分類されるものの多くは日本の一地方、または一地域で使用される言語変種を指し示している可
能性が高いが、それ以外の解釈も可能であるため、念のため、A とは別扱いにした。

F 特定の位相や人々が前項要素となっているもの： 8 表現
ヤンキー・百姓・日本人・外国人・黒人・廓・西島・軍隊

G その他①・日本の地方・地域を表している可能性が高いが、別の解釈もできる語が前項要素となっているもの： 3 表現
雪国・他国・中国

H その他②・特定の位相や人々を表している可能性が高いが、地域を指し示している可能性も考えられる語が前項要素となっているもの： 3 表現
コックニー・下町・里¹⁶

ここで見いだした言語変種を表す複合語「～訛(り)」に関する簡単な考察は、3.3 で行う。

3.2 国立国語研究所が構築し、公開している各種コーパスを使用した調査

3.1 では「NDL Ngram Viewer」を利用して、日本語で使用されている言語変種を表す複合語「～訛(り)」を見いだした。ただ、注 12 でも触れているとおり、「NDL Ngram Viewer」には挙動に不明瞭な点もある。また、岡田(2024, 印刷中)でもまとめているとおり、「NDL Ngram Viewer」で検索できる全文テキストデータは膨大であるものの、4 節で扱う「国立国会図書館デジタルコレクション」の全文テキストデータと比較すると限定されている。その結果、いわゆる「検索漏れ」が生じている可能性もある。

そこで、ここでは、補足的に国立国語研究所が構築し、公開している各種コーパスにおいて、どのような言語変種を表す複合語「～訛(り)」が存在しているのか、確認した。具体的には、コーパス検索アプリケーション『中納言』で利用できる「まとめて検索『KOTONOHA』」で、以下のような条件を設定し、各種コーパスを検索した。

(d) キー： 「品詞」の「大分類」が「名詞」
後方共起「キーから 1 語」： 語彙素「訛(り)」／「品詞」の「大分類」が「名詞」

その結果、各コーパスにおける検索件数は以下ようになった。

- ・『現代日本語書き言葉均衡コーパス 第 1 部』(BCCWJ1)： 169 件
- ・『現代日本語書き言葉均衡コーパス 第 2 部』(BCCWJ2)： 29 件
- ・『日本語話し言葉コーパス』(CSJ)： 16 件
- ・『日本語日常会話コーパス』(CEJC)： 2 件
- ・『昭和話し言葉コーパス』(SSC)： 1 件
- ・『名大会話コーパス』(NUCC)： 3 件
- ・『日本語歴史コーパス』(CHJ)： 23 件
- ・『昭和・平成書き言葉コーパス』(SHC)： 80 件
- ・『日本語諸方言コーパス』(COJADS)： 5 件

¹⁶ 「里訛り」については、注 10, 注 11 も参照のこと。

以上の検索結果の中から目視で言語変種を表す複合語「～訛(り)」を抽出したのち、3.1で示した分類 A から H に含まれないものを整理した結果が、以下の A', B', D', E', F', H'である。なお、カッコ内は、当該複合語が観察されたコーパスと用例数を示している。また、3.1 での「NDL Ngram Viewer」の検索結果の整理に倣い、コーパスでの検索結果の整理は(「語彙素」ではなく)「書字形」を利用したため、ここでも異なり書字形数の単位は「表現」となる。また、後項要素「訛(り)」を省略して列記しているのは、これまでと同様である。

A' 日本の特定の地方・地域・都市などが前項要素となっているもの： 7 表現

下関伊 (SHC・1)・柏崎 (CHJ・1)・下呂 (BCCWJ1・1)・久留米 (CHJ・1)・諫早 (SHC・1)・我那覇 (BCCWJ2・1)・河間 (COJADS・1)

B' 外国の特定の国・地方・地域・都市、あるいは民族などが前項要素となっているもの：

36 表現

イングランド (BCCWJ1・2)・グラスゴー (BCCWJ1・1)・ランカシャー (CHJ・1)・プロヴァンス (BCCWJ1・1)・ミュンヘン (CHJ・1, SHC・1)・ナポリ (BCCWJ1・1, NUCC・1)・ギリシア (BCCWJ1・1)・ハンガリー (BCCWJ1・1)・スロヴェニア (BCCWJ1・1)・ブルガリア (BCCWJ1・1)・レヴァント (BCCWJ1・1)・アラブ (BCCWJ1・1)・マダガスカル (BCCWJ1・1)・南アフリカ (BCCWJ2・1)・シカゴ (BCCWJ1・2)・テネシー (BCCWJ1・1)・モンタナ (BCCWJ2・1)・ルイジアナ (BCCWJ1・1)・ジャマイカ (BCCWJ1・1)・タイ (BCCWJ1・1)・カンボジア (BCCWJ1・1)・フィリピン (BCCWJ1・2, BCCWJ2・1, CSJ・1)・ジャカルタ (BCCWJ1・1)・オーギー (BCCWJ1・1)・ニュージーランド (CSJ・1)・慶尚道 (BCCWJ1・3, BCCWJ2・2, SHC・1)・全羅道 (BCCWJ1・2)・釜山 (BCCWJ2・2)・プサン (BCCWJ1・1)・北朝鮮 (BCCWJ1・1)・平安道 (BCCWJ1・2)・四川 (BCCWJ1・1)・香港 (BCCWJ1・2)・上海 (BCCWJ1・1)・浙江省 (BCCWJ1・1)・山西省 (CHJ・1)

D' 地方・地域は特定しないが、主に地方部(都市部以外)を意味する語が前項要素となっているもの： 1 表現

地元 (BCCWJ1・2)

E' 言語名が前項要素となっているもの： 7 表現

イタリア語 (BCCWJ1・1, BCCWJ2・1)・ロシア語 (BCCWJ2・1)・イディッシュ (BCCWJ1・1)・タイ語 (BCCWJ2・1)・ベトナム語 (BCCWJ1・2)・朝鮮語 (BCCWJ1・1)・母語 (CSJ・1)

F' 特定の位相や人々が前項要素となっているもの： 7 表現

べらんめえ (SHC・1)・江戸子 (SHC・1)・上流 (BCCWJ1・1)・武士 (BCCWJ1・1)・標準 (CSJ・1)・BBC (CSJ・1)・NHK (CSJ・1)

H' その他②・特定の位相や人々を表している可能性が高いが、地域を指し示している可能性も考えられる語が前項要素となっているもの： 1 表現

コックニ (BCCWJ1・1)

上にまとめた表現を「国立国会図書館デジタルコレクション」（詳細は4節を参照）で検索したところ、「コックニ」「下呂」「河間」「我那覇」以外は用例が確認できた。この結果は、「NDL 全文テキストデータ」には存在しているにもかかわらず、「NDL Ngram Viewer」ではヒットしないものもあるということを示している。また、各種コーパスでも多様な言語変種を表す複合語「～訛(り)」の例を得ることができるが、それぞれの書字形の出現回数は少なく、多くは1例にとどまる。このことから、これらのコーパスで使用実態を観察するには十分な用例が確保できないことも分かる。

ここで見いだした言語変種を表す複合語「～訛(り)」に関する簡単な考察は、3.1 で報告した結果と合わせて、3.3 で行う。

3.3 本節のまとめ

3 節では、「NDL Ngram Viewer」に加え国立国語研究所が構築し、公開している各種コーパスを利用し、日本語で使用されている言語変種を表す複合語「～訛(り)」にはどのようなものが存在するかを確認した。その結果、分類 A から H および該当するプライム付き項目に列記したように、多様な前項要素を持つ表現が確認された。これは、2 節でまとめた国語辞典の立項状況に関する調査からはいかがい知ることができなかったことである。このことから、言語変種を表す複合語「～訛(り)」は一定の生産性を持つと考えられる。

分類 A から H を見ていると、言語変種を表す複合語「～訛(り)」は、地域変種を指すもの (A, C, D など) と社会変種 (F など) を指すものとに分けることができそうである。そのうえで、地域変種を指すものと比較して社会変種を指すものが少ないことも注目に値する。これは、「現代の日本の社会では、江戸時代の士農工商のような身分制度がなく、国民の多くが中流層に位置づけられ、はっきりとした階級区分は存在しないと言え」ることから、「方言といえば、地域方言しか思い浮かばない人も多いだろう」(川村 2015) という現状を反映していると考えられる。また、地域変種を指すものにはいわゆる旧国名やその別称が使用されていることが多いことにも気づく。これは、同一都道府県内にも複数の地域変種が存在しているため、都道府県名を使用して地域変種を指し示すことは適当ではない場合も多いことを反映している可能性も考えられる。

なお、分類 A から H にまとめたものの中には、現在では使用に慎重さを要する語を前項要素とするものもある。このことから、「言葉や発音がくずれる」という語義を持つ「～訛(り)」は、侮蔑的なコノテーションを有する語と結びつきやすい可能性もある。ただし、現在では使用に慎重さを要する語であっても、当該用例が生み出された時代においては侮蔑的なコノテーションを有していなかった可能性もある。この点の考察も今後の課題としたい。

最後に、これらの複合語には、言語変種そのものを指す用法と、中間言語的な変種を指す用法とがあることを指摘したい。たとえば、「越後訛(り)」という複合語の場合、「越後地方(新潟県)独特の発音のしかた」(『日国』の語釈より)という越後地方で話されている言語変種そのものを指す。一方、「越後訛(り)」は越後地方で話されている言語変種そのものを指すのではなく、「日本語」(この文脈の場合は「標準語」あるいは「共通語」)

に越後地方で話されている言語変種の要素が混ざる状態、いわば中間言語的な変種を指す用例も散見されたということである。このような用法は、「越後訛りの日本語」といった表現から生まれたとも考えられる。そして、このような中間言語的な変種を指す用法は、分類 B, D, E など複数、観察された。特に E でその傾向が顕著である。しかし、今回の調査ではこのような用法がどの程度観察されるのか、その詳細を整理するまでには至らなかった。そのような状態での指摘ではあるが、2 節で確認した二つの国語辞典には、言語変種を表す複合語「～訛(り)」にこのような中間言語的な変種を指す用法もあることが必ずしも明示されていなかった¹⁷ため、ここではこのような用法が存在することに言及した次第である。

4. 日本語で使用されている言語変種を表す複合語「～訛(り)」の資料上の出現状況の一側面―「国立国会図書館デジタルコレクション」を利用した素描―

3 節では、「NDL Ngram Viewer」と国立国語研究所が構築し、公開している各種コーパスを利用し、日本語で使用されている言語変種を表す複合語「～訛(り)」にどのようなものがあるのかを整理した。ただ、その結果は、あくまで「どのようなものがあるのか」を示したにすぎず、どの複合語が多く使われているのか、また、使用される語に経年変化が観察されるかなど、「どのように使用されているのか」という点は不明である。そこで、ここでは、「国立国会図書館デジタルコレクション」（以下、「デジコレ」と表記）を利用し、日本語で使用されている言語変種を表す複合語「～訛(り)」の資料上の出現状況の一端を素描することを試みる。

岡田（2024，印刷中）でも触れているとおり、「デジコレ」は「NDL 全文テキストデータ」の検索ができるプラットフォームの一つである。3.1 で使用した「NDL Ngram Viewer」とは異なり単純な文字列検索しかできないが、検索できるデータ量が圧倒的に多く、資料の種類も豊富である（2026 年 7 月末時点で、図書約 248 万点、雑誌約 141 万点、博士論文 18 万点、官報約 2 万点などが検索可能）。また、出版年月日や NDC 分類といった情報を活用した集計も可能である。

ここでは、3 節で見いだした言語変種を表す複合語「～訛(り)」を「デジコレ」で検索し、その検索結果件数をもとに、それぞれの語の資料上の出現状況の一側面を素描する。今回の検索で指定した検索語（文字列）は以下のとおりである。なお、「○○」は 3 節で見いだした言語変種を表す複合語「～訛(り)」の前項要素を意味する。

(e) ○○訛 -○○訛音 -○○訛言 -○○訛語

（「NDL Ngram Viewer」とは異なり）「デジコレ」での検索では、「○○訛」で検索することで「○○訛(り)」も漏れなくヒットするようであったため、「○○訛」のみを検索ことにした。一方、「○○訛」で検索をすると「○○訛音」「○○訛言」「○○訛語」という語（文字列）もヒットしてしまう場合が散見された。そこで、検索に当たっては一律に

¹⁷ ただ、『日国』の「田舎訛り」の語釈、すなわち「言葉に田舎のなまりがあること。また、標準語と違った発音、アクセント、ことばぐせ。国なまり。田舎調子。」は、「標準語」に「田舎のなまり」が観察されることを意味しているように解釈でき、これは、発表者のいう中間言語的な変種を指す語釈と見なせるかもしれない。

「〇〇訛音」「〇〇訛言」「〇〇訛語」を NOT 検索の条件として設定した。ただ、このような工夫を施したとはいえ、検索結果には発表者が意図しない検索上の「ノイズ」は含まれていると思われる。しかし、単純な文字列検索しかできない「デジコレ」では、完全に検索上の「ノイズ」を除去することは不可能である。したがって、検索結果件数を解釈する際には、検索上の「ノイズ」も含まれていることを念頭に置く必要があることを、あらかじめ申し添える。

また、「デジコレ」での検索結果は、検索語（文字列）の出現頻度ではなく、検索語（文字列）が含まれる資料の点数であることにも留意する必要がある。

なお、今回の検索では、3 節で見いだした言語変種を表す複合語「～訛（り）」のうち、前項要素が 1 文字のもの（「京」「国」など）は検索対象から外している。これは、前項要素が 1 文字のものは検索上の「ノイズ」が多種多様だったため、NOT 検索を活用してもそれらを除去することが難しいと判断したことによる。また、前項要素が「NHK」「BBC」は「デジコレ」においては(e)で示した検索条件ではうまく検索できなかった¹⁸。そのため、こちらも今回の検索対象からは除外している。

4.1 概観

言語変種を表す複合語「～訛（り）」のうち、そもそも、分類 A から H のうち、検索結果件数が多いカテゴリはいったいどれなのであろうか。この点を確認すべく、3 節で見いだした言語変種を表す複合語「～訛（り）」を「デジコレ」で検索し、分類ごとに検索結果件数を合計した。その結果が、以下の表 1 である。

表 1 「国立国会図書館デジタルコレクション」に見る

言語変種を表す複合語「～訛（り）」の前項要素： 分類別の検索結果件数と構成比

前項要素の種類	検索結果件数	構成比
A 日本の特定の地方・地域・都市など	50701	44.7%
B 外国の特定の地方・地域・都市、あるいは民族など	26383	23.3%
C 方角・方向などが含まれる表現	6651	5.9%
D 地方・地域は特定しないが、主に地方（都市部以外）を意味する語	23028	20.3%
E 言語名	4046	3.6%
F 特定の位相や人々	932	0.8%
G その他①	813	0.7%
H その他②	835	0.7%

表 1 からは、言語変種を表す複合語「～訛（り）」の前項要素としては、分類 A、すなわち日本の地域変種を表現する場合に多く使用されていることが読み取れる。また、検索結果件数の観点からは D も一定数、使用されていることが分かる。すなわち、言語変種を表す複合語「～訛（り）」は、多くの場合、特定地域を明示するか否かはともかく、地域変種を表現することが多いといつてよさそう。

一方、分類 B に分類されるものも多く使用されていることが分かる。こちらは、地域変種を指す場合も含まれているだろうが、「ドイツ訛りの英語」のような中間言語的な変種を指す場合も含まれていると考えられる。ただ、検索結果件数が非常に多く、今回の調査では

¹⁸ 「NHK 訛」「BBC 訛」で検索をすると、「訛」を含まない「NHK」「BBC」という文字列のみを含む資料まで多数ヒットしてしまうようである。一方、「NHK 訛り」「BBC 訛り」で検索をした場合は発表者が意図したものがヒットするが、件数は僅少であった（前者が 2 件、後者が 6 件）。

用法に関する詳細な検討までは至らなかった。今後の課題としたい。

また、分類 F の検索結果件数が少ないことに注目したい。3.3 で、前項要素の種類の観点から、地域変種を指すものと比較して社会変種を指すものが少ない傾向にあることを指摘したが、この傾向は、検索結果件数の場合も同様であるといえる。

表 1 からは、言語変種を表す複合語「～訛(り)」は、前項要素として主に分類 A, B, D に属するものが使用されることが多いと分かる。では、具体的にどのような前項要素が多く使用されているのであろうか。この点を明らかにすべく、検索結果件数が 1000 以上の前項要素を件数が多い順に並べてみた。その結果が、以下の表 2 である。

表 2 「国立国会図書館デジタルコレクション」に見る

言語変種を表す複合語「～訛(り)」の前項要素： 検索結果件数 1000 以上（降順表示）

前項要素	件数	前項要素	件数	前項要素	件数	前項要素	件数	前項要素	件数
お国	7941	南部	3161	アメリカ	1701	江戸	1388	秋田	1206
東北	7624	九州	2384	上方	1666	ロンドン	1374	方言	1173
田舎	5017	ドイツ	2361	アイルランド	1590	スコットランド	1341	京都	1074
関西	4944	土地	2026	薩摩	1556	越後	1330	津軽	1063
地方的	4036	外国	1715	大阪	1484	フランス	1277	土佐	1039

表 2 を見ると、「お国訛(り)」「東北訛(り)」が圧倒的に多く使用される傾向にあることが分かる。また、分類 D のように地方・地域を特定しない表現（「お国」「田舎」「地方的」「土地」）よりも、地方・地域・国などを特定する表現が多く使用される傾向にある。

なお、表 2 に示している 25 表現のうち、『日国』や『大辞泉』に立項されているのは 9 表現であった。このことから、日本語で多く使用されているものは国語辞典にも立項される傾向にあると考えられる。ただ、「ドイツ」「アメリカ」のような外国の国名や地域・地方・都市名は、「デジコレ」での検索結果件数は多いものの『日国』や『大辞泉』には立項されていない。これは、言語変種を表す「～訛(り)」は一定の生産性があることから、前項要素に様々な国名・地名・地域名が入ったとしても類推でどのような言語変種を指しているのか推測でき、わざわざ国語辞典に立項する必要性が低いためだとも考えられる。

以下、検索結果件数が特に多かった分類 A, B, D について、さらにその内実を観察することにする。

4.2 日本の特定の地方・地域・都市などが前項要素となっているものについて

3 節で確認した分類 A と A' のうち、1 文字である「京」と「東」、および「デジコレ」では用例が確認できなかった「下呂」「河間」「我那覇」を除いた 126 表現について、「デジコレ」を使用し、検索結果件数を調べたうえで累積比率を求めた。その結果が、次ページの表 3 である。

表 3 からは、ここに分類されるものは種類こそ多いものの、多く使用されるものは限定的であることが分かる。126 表現のうち上位 11 表現、すなわち上位約 9% で累積比率が過半数となる。特に「東北」と「関西」だけで累積比率が約 25% となる。また、四分位数を利用し「外れ値」¹⁹を求め、外れ値に該当する検索結果件数を有するものを確認したところ、「東

¹⁹ ここでは、第 3 四分位数 + 四分位範囲 × 1.5 以上のもの。

表3 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率①

A 日本の特定の地方・地域・都市などが前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率	前項要素	件数	累積比率
東北	7624	15.0%	坂東	299	78.7%	下総	99	95.2%
関西	4944	24.8%	三河	296	79.3%	美濃	98	95.4%
九州	2384	29.5%	難波・飛騨	293	79.8%	遠州	97	95.6%
上方	1666	32.8%	岡山	290	81.0%	伊豆・甲斐	95	95.8%
薩摩	1556	35.8%	福島	287	81.5%	上総	94	96.1%
大阪	1484	38.8%	水戸	279	82.1%	山口・天草	91	96.3%
江戸	1388	41.5%	能登	275	82.6%	安房	87	96.7%
越後	1330	44.1%	新潟	264	83.2%	唐津	86	96.9%
秋田	1206	46.5%	博多	258	83.7%	福井	84	97.0%
京都	1074	48.6%	加賀・出羽	245	84.2%	福岡	83	97.2%
津軽	1063	50.7%	甲州	243	85.1%	播州	80	97.3%
土佐	1039	52.8%	常陸・岩手	227	85.6%	伊賀・肥前	76	97.5%
関東	998	54.7%	北海道	224	86.4%	備前	75	97.8%
奥州	883	56.5%	阿波	191	86.8%	相模	70	97.9%
出雲	791	58.0%	青森	180	87.2%	但馬	68	98.1%
広島	665	59.4%	鎌倉・佐賀・筑後	178	87.5%	伊那・武州・房州	67	98.2%
日本	601	60.5%	北陸	176	88.6%	埼玉	66	98.6%
名古屋	595	61.7%	越中	174	88.9%	浪花	64	98.7%
四国	549	62.8%	沖縄	173	89.3%	盛岡	60	98.8%
上州	546	63.9%	信濃	171	89.6%	秩父	59	98.9%
伊予	536	64.9%	富山・豊後	163	89.9%	陸奥	58	99.1%
東京	532	66.0%	蝦夷	162	90.6%	羽後	54	99.2%
仙台	527	67.0%	河内	161	90.9%	桑名	50	99.3%
会津	522	68.1%	近江	143	91.2%	山陰・丹後	44	99.4%
長崎	490	69.0%	金沢	142	91.4%	島原	42	99.5%
肥後	472	69.9%	丹波	130	91.7%	志摩	40	99.6%
茨城	439	70.8%	讃岐	128	92.0%	淡路	39	99.7%
紀州	427	71.7%	日向	123	92.2%	久留米	34	99.7%
山形	417	72.5%	佐渡・松山	122	92.4%	函館	33	99.8%
みちのく	398	73.3%	越前	119	92.9%	紀伊	25	99.9%
栃木	385	74.0%	大分	118	93.1%	佐久	23	99.9%
鹿児島	380	74.8%	筑紫	113	93.4%	吉備	20	99.9%
熊本	375	75.5%	神戸	112	93.6%	日立	11	100.0%
信州	352	76.2%	若狭	105	93.8%	諫早	9	100.0%
伊勢	346	76.9%	江州・大和	104	94.0%	柏崎	6	100.0%
長州	302	77.5%	高知・木曾・庄内	102	94.4%	下関伊	3	100.0%
尾張	300	78.1%	駿河	101	95.0%			

北」から「奥州」までが、それに該当した。日本の特定の地方・地域・都市などを前項要素として持つ言語変種を表す「～訛(り)」として特によく使用される表現と見なせるかについては俄かに結論が出せないが、多く使用されるものは限定的であるといつてよからう。なお、表3の中央値は163である。

また、いわゆる旧国名やその別称が、対応する現行の都道府県名よりも好まれて使用される傾向が見て取れる。たとえば、「越後」は1330件であるのに対し、「新潟」は264件であり、前者は後者の約5.0倍に達する。同様に、「薩摩」1556件は「鹿児島」380件の約4.1倍、「土佐」1039件は「高知」102件の約10.2倍の検索結果件数が存在する。また、「上州」546件、「信州」352件、「信濃」171件なのに対し、「群馬」や「長野」は「NDL Ngram

Viewer」や国立国語研究所が構築、公開している各種コーパスによる検索では抽出されず、今回の集計対象には含まれていない²⁰。3.3 で、前項要素の種類の観点から、いわゆる旧国名が多く使用される傾向を指摘したが、この傾向は、検索結果件数の場合も同様であるといえる。

4.3 外国の特定の国・地方・地域・都市、あるいは民族などが前項要素となっているもの

3 節で確認した分類 B と B' の 97 表現について、「デジコレ」を使用し、検索結果件数を調べたうえで累積比率を求めた。その結果が、次ページの表 4 である。

表 4 からは、ここに分類されるものは種類こそ多いものの、多く使用されるものは限定的であることが分かる。97 表現のうち上位 10 表現、すなわち上位約 10% で累積比率が過半数となる。また、4.2 で用いたのと同様の手法で「外れ値」を求め、外れ値に該当する検索結果件数を有するものを確認したところ、「ドイツ」から「スペイン」までが、それに該当するということが分かった。外国の特定の国・地方・地域・都市などを前項要素として持つ言語変種を表す「～訛(り)」として特によく使用される表現と見なせるかについても俄かに結論が出せないが、分類 A と同様、多く使用されるものは限定的であるといっていよかろう。なお、表 4 の中央値は 105 であった。

また、表 4 の上位には「外国」と非常に漠然とした前項要素も存在するが、これは日本語話者にとって耳馴染みのない発音が「外国訛(り)」と表現されることが多いためではないだろうか。一方、「アメリカ」「アイルランド」「ロンドン」「スコットランド」「テキサス」といった英語が通用する国・地域・都市名も散見される。このことは、日本語の文章においても「アメリカ訛(り)」「アイルランド訛(り)」といった形で英語の変種に言及されることが少なくないという事情を反映しているとも考えられる。あわせて、ドイツ語が通用する国・地域・都市名（「ドイツ」「ザクセン」「ベルリン」など）も少なくない検索結果件数が認められることにも注目したい。

²⁰ ここでの考察の段階で「デジコレ」を確認したところ、「群馬訛(り)」は 26 件、「長野訛(り)」は 9 件、それぞれ確認できた。つまり、「群馬訛(り)」や「長野訛(り)」は「NDL 全文テキストデータ」に含まれているにもかかわらず「NDL Ngram Viewer」の検索ではヒットしなかった事例ということになる。

表4 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率②

B 外国の特定の国・地方・地域・都市, あるいは民族などが前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率
ドイツ	2361	8.9%	ハンガリー	134	88.9%
外国	1715	15.4%	ウクライナ	130	89.4%
アメリカ	1701	21.9%	コルシカ	122	89.9%
アイルランド	1590	27.9%	四川	120	90.3%
ロンドン	1374	33.1%	アルザス・倫敦	118	90.8%
スコットランド	1341	38.2%	スラヴ	110	91.6%
フランス	1277	43.1%	フィリピン	100	92.0%
イタリア	931	46.6%	ジャマイカ・プロヴァンス	99	92.4%
英国	930	50.1%	ナポリ	96	93.1%
テキサス	833	53.3%	仏蘭西	90	93.5%
イギリス	793	56.3%	浙江	88	93.8%
スペイン	632	58.7%	伊太利・アラブ・慶尚道	84	94.1%
オーストラリア	576	60.8%	グラスゴー	82	95.1%
ブルックリン	525	62.8%	山西	78	95.4%
ロシア	461	64.6%	イタリー	75	95.6%
ボストン	426	66.2%	北京	72	95.9%
ニューヨーク	420	67.8%	支那・蘇州・ランカシャー	69	96.2%
ウィーン	412	69.4%	ギリシア	66	96.9%
朝鮮	309	70.5%	シュワーベン	63	97.2%
オランダ	298	71.7%	テネシー	57	97.4%
スイス	279	72.7%	羅馬	48	97.6%
インド	270	73.7%	瑞西・天津	47	97.8%
コクニー	265	74.7%	平安道	45	98.1%
ザクセン	259	75.7%	ニュージーランド	44	98.3%
ポーランド	258	76.7%	ガスコン・ペリゴール	43	98.4%
ベルリン	250	77.6%	巴里・全羅道	36	98.7%
ユダヤ	245	78.6%	南アフリカ	35	99.0%
イングランド	238	79.5%	ミュンヘン	34	99.1%
山東	233	80.4%	シカゴ	29	99.2%
独逸	231	81.2%	北朝鮮	28	99.4%
パリ	204	82.0%	ルイジアナ	27	99.5%
タイ	201	82.8%	エディンバラ	26	99.6%
ヨークシア	191	83.5%	香港	25	99.7%
広東	190	84.2%	京城	20	99.7%
米国	181	84.9%	オーギー・釜山	12	99.8%
バイエルン	174	85.6%	モンタナ	10	99.9%
ニューイングランド	168	86.2%	カンボジア・山西省	7	99.9%
湖南	148	86.8%	ジャカルタ	6	99.9%
上海	148	87.3%	ブルガリア	5	100.0%
ラテン	144	87.9%	浙江省	4	100.0%
ガスコーニュ	140	88.4%	スロヴェニア・レヴァント	3	100.0%
			イッディシュ・ブサン・マダガスカル	1	100.0%

4.4 地方・地域は特定しないが、主に地方部（都市部以外）を意味する語が前項要素となっているものについて

3 節で確認した分類 D と D' のうち、1 文字である「国」「浜」を除いた 17 表現について、「デジコレ」を使用し、検索結果件数を調べたうえで累積比率を求めた。その結果が、次ページの表 5 である。

表 5 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率③

D 地方・地域は特定しないが、主に地方部を意味する語が前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率	前項要素	件数	累積比率
お国	7941	34.7%	ふるさと	572	94.3%	郷土	119	98.7%
田舎	5017	56.6%	御国	400	96.1%	郷里	94	99.1%
地方的	4036	74.3%	おくに	207	97.0%	植民地	84	99.5%
土地	2026	83.1%	故里	144	97.6%	ふる里	65	99.8%
方言	1173	88.2%	古里	143	97.6%	地元	54	100.0%
故郷	817	91.8%	地方	136	98.2%			

表 5 からは、分類 D については特に「お国訛（り）」に検索結果件数が集中する傾向が読み取れる。別表記のため表 5 では別扱いになっているが「御国」「おくに」、さらに今回は漢字 1 字であるため検索から除外したが、「国」まで含めるとその傾向はより顕著になると思われる。このことから、「（お）国訛（り）」が定型的な表現として定着していることをうかがわせる。

一方、「田舎訛（り）」も検索結果件数が多いことに注目したい。「田舎」に対して『大辞泉』では以下の四つの語釈を与えているが、語釈 3 は表 5 の「故郷」「ふるさと」「故里」「古里」「郷里」「ふる里」と同義と見なしてよからう。そのように考えると、分類 D は「（お）国訛（り）」とあわせ、故郷の言語変種を表す表現が大半を占めていると考えてもよさそうである。

ただし、以下に示した『大辞泉』の特に語釈 1 や語釈 4 を踏まえると、「田舎訛（り）」の検索結果には「都会的ではない」「洗練されていない」言語変種を指し示す表現が一定数含まれていることも予想される。

- ・ 1 都会から離れた地方。
- 2 田畑が多く、のどかな所。人家が少なく、静かでへんぴな所。
- 3 生まれ故郷。郷里。父母や祖父母のふるさとについてもいう。
- 4 （名詞に付き、接頭語的に用いて）素朴・粗暴・やぼなどの意を表す。

4.5 資料： 分類 C, E, F, G, H について

4.1 での議論を踏まえ、4.2 から 4.4 まで、検索結果件数が特に多かった分類 A, B, D について、さらにその内実を観察してきた。その他の分類についても仔細を確認したいのであるが、発表時間の関係もあり詳細に扱うことができない。そこで、ここでは、これまでに触れることができなかった分類 C, E, F, G, H について、「デジコレ」での検索結果件数と累積比率をまとめた結果を資料として提示する。それぞれの分類においてどのような前項

要素が使用されやすいか、興味を持たれた方は以下の表をご確認いただきたい。

表 6 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率④
C 方角・方向などが含まれる表現が前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率
南部	3161	47.5%	北部	253	88.7%
東国	738	58.6%	中西部	252	92.5%
北国	685	68.9%	南方	242	96.1%
西部	539	77.0%	北方	159	98.5%
西国	522	84.9%	南国	100	100.0%

表 7 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率⑤
E 言語名が前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率	前項要素	件数	累積比率
英語	751	18.6%	外国語	298	84.8%	漢語	40	97.9%
ドイツ語	729	36.6%	中国語	157	88.7%	母国語	34	98.8%
フランス語	684	53.5%	ロシア語	141	92.2%	タイ語	30	99.5%
スペイン語	497	65.8%	イタリア語	138	95.6%	ベトナム語	10	99.8%
日本語	472	77.4%	朝鮮語	56	97.0%	母語	9	100.0%

表 8 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率⑥
F 特定の位相や人々が前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率	前項要素	件数	累積比率
黒人	278	29.8%	日本人	82	85.3%	江戸子	16	97.5%
外国人	204	51.7%	軍隊	51	90.8%	べらんめえ	11	98.7%
百姓	127	65.3%	西島	28	93.8%	上流	8	99.6%
ヤンキー	104	76.5%	武士	19	95.8%	標準	4	100.0%

表 9 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率⑦
G その他①・日本の地方・地域を表している可能性が高いが、
別の解釈もできる語が前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率	前項要素	件数	累積比率
中国	650	80.0%	他国	126	95.4%	雪国	37	100.0%

表 10 「国立国会図書館デジタルコレクション」の検索結果件数と累積比率⑧
H その他②・特定の位相や人々を表している可能性が高いが、
地域を指し示している可能性も考えられる語が前項要素となっているもの

前項要素	件数	累積比率	前項要素	件数	累積比率
下町	570	68.3%	コックニー	265	100.0%

5. おわりに

本発表では、「NDL 全文テキストデータ」を主な調査対象とし、言語変種を表す複合語「～

訛(り)」にはどのようなものが存在し、また、それらがどのように使用されているのか、その諸相について素描を試みた。その結果、本調査の対象とした国語辞典には立項されていない多様な複合語が使用されており、この形式が一定の生産性を有する可能性を示した。また、言語変種を表す「～訛(り)」には地域変種を指すものが多く、社会変種を指すものは相対的に少なかった。そのうえで、中間言語的な変種を指す用法も確認できた。さらに、言語変種を表す「～訛(り)」の使用は一部の分類に集中していること、とりわけ検索結果件数の多い分類においては多く使用される表現は一部に限定されることが確認された。

ただ、本発表はどのような複合語がどの程度多くの資料に現れるのかという議論に終始してしまい、使用実態を経年変化、媒体、ジャンルといった観点から整理することまでには至らなかった。ただし、岡田(印刷中)でも触れているとおり、こちらは、「デジコレ」の検索結果に表示される「ファセット項目」を活用すれば簡便に確認できる。こちらは発表中にデモンストレーション的にお見せする予定である。

なお、1節でも触れたとおり、本発表は「NDL 全文テキストデータ」を日本語研究に利用する方法を実践した事例報告としての意味合いもあったが、岡田(印刷中)で提案した「NDL Ngram Viewer」と「デジコレ」を併用した手法で、一定の成果を上げることができると考える。もっとも、本発表で示した検索結果件数は、検索対象とした文字列の出現回数ではなく、その文字列を含む資料の点数であり、検索上のノイズも含まれている。また、今回の作業を通じて、「NDL Ngram Viewer」の正規表現を利用した検索には、検索漏れが生じることも分かった。この点も本発表の成果の一つとして位置づけられよう。

最後に残された大きな問題点として、それぞれの用法や使用文脈を詳細に検討するまでには至らなかったことがあげられる。これは、今回の検索結果件数が膨大であり、得られた用例を逐一確認することが不可能だったためである。そのため、本文中では断言を避ける表現が多くなっている。今回に限らず、「NDL 全文テキストデータ」を利用した調査では、石井(2014)のいう「コーパスのパラドクス」ともいえる逆説的な事態、つまり「コーパスが大きくなればなるほど、そこから引き出されるデータも多くなり、分析者にとって手に負えないものになってしまう」事態に必ず直面する。ただ、発表者は今なお、有効な解決策を見いだせていない(岡田 2024, 印刷中も参照)。

謝 辞

本稿は JSPS 科研費 JP25K04063, JP25K04144 の助成を受けたものである。

文 献

- 石井正彦(2014). 「コーパスを用いた語彙研究の特徴」 田野村忠温 [編] コーパスと日本語学 講座日本語コーパス 6, pp.3-13, 朝倉書店.
- 岡田祥平 (2024). 「「国立国会図書館デジタルテキストデータ」を利用した日本語研究の可能性と限界」 日本語学会 2024 年度春季大会予稿集, pp.143-148. https://www.jpling.gr.jp/wp-content/uploads/2024/08/2024a_all-2.pdf
- 岡田祥平 (印刷中). 「「国立国会図書館全文テキストデータ」を利用した近・現代日本語研究ことはじめ—「国立国会図書館デジタルコレクション」と「NDL Ngram Viewer」の活用法に関する覚書—」 新潟大学教育学部研究紀要 人文・社会科学編, 19:1.
- 川村陽子(2015). 「IV 社会方言 1 社会階級と言語」 田中春美・田中幸子 [編著] よくわかる社会言語学, pp.38-39, ミネルヴァ書房.

- 小木曾智信・近藤明日子・高橋雄太・間淵洋子 [編] (2024). 『『昭和・平成書き言葉コーパス』の設計・構築・公開』情報処理学会論文誌, 65:2, pp.278-291.
<https://doi.org/10.20729/00232292>
- 小磯花絵・天谷晴香・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香 (2023). 『『日本語日常会話コーパス』設計と構築』国立国語研究所論集, 24, pp.153-168. <https://doi.org/10.15084/00003692>
- 国立国語研究所 (2006). 『日本語話し言葉コーパスの構築法』国立国語研究所.
<https://doi.org/10.15084/00001357>
- 藤村逸子・大曾美恵子・大島ディヴィッド義和 (2011). 『会話コーパスの構築によるコミュニケーション研究』藤村逸子・滝沢直宏 [編] 言語研究の技法：データの収集と分析, pp. 43-72, ひつじ書房.
- 前川喜久雄 [監修] 山崎誠 [編] (2014). 『書き言葉コーパス—設計と構築— 講座日本語コーパス 2』朝倉書店.
- 丸山岳彦・小磯花絵・西川賢哉 (2022). 『『昭和話し言葉コーパス』の設計と構築』国立国語研究所論集, 22, pp.197-221. <https://doi.org/10.15084/00003522>

利用データ・関連 URL

- コーパス検索アプリケーション『中納言』 <https://chunagon.ninjal.ac.jp/>
- 国立国語研究所 (2025). 『現代日本語書き言葉均衡コーパス』(中納言バージョン 2.7.3, データバージョン 形態論情報 2021.03, 分類語彙表情報 2025.03)
<https://clrd.ninjal.ac.jp/bccwj/> (2026 年 8 月 7 日確認)
- 国立国語研究所 (2026). 『現代日本語書き言葉均衡コーパス 第 2 部』(中納言バージョン 2.7.3, バージョン 2026.03) <https://www2.ninjal.ac.jp/BCCWJ2/> (2026 年 8 月 7 日確認)
- 国立国語研究所 (2018). 『日本語話し言葉コーパス』(中納言バージョン 2.7.2, バージョン 2018.01) <https://clrd.ninjal.ac.jp/cs/> (2026 年 8 月 7 日確認)
- 国立国語研究所 (2023). 『日本語日常会話コーパス』(中納言バージョン 2.7.2, バージョン 2023.03) <https://www2.ninjal.ac.jp/conversation/cejc.html> (2026 年 8 月 7 日確認)
- 国立国語研究所 (2022). 『昭和話し言葉コーパス』(中納言バージョン 2.7.2, バージョン 2022.02) <https://www2.ninjal.ac.jp/conversation/showaCorpus/> (2026 年 8 月 7 日確認)
- 国立国語研究所 (2022). 『名大会話コーパス』(中納言バージョン 2.7.2, バージョン 2022.10) <https://mmsrv.ninjal.ac.jp/nucc/> (2026 年 8 月 7 日確認)
- 国立国語研究所 (2026). 『日本語歴史コーパス』(中納言バージョン 2.7.2, バージョン 2026.3) <https://clrd.ninjal.ac.jp/chj/> (2026 年 8 月 7 日確認)
- 小木曾智信・近藤明日子・高橋雄太・田中牧郎・間淵洋子編 (2023) 『昭和・平成書き言葉コーパス』(中納言バージョン 2.7.2, バージョン 2023.5) <https://clrd.ninjal.ac.jp/SHC/> (2026 年 8 月 7 日確認)
- 国立国語研究所 (2026). 『日本語諸方言コーパス』(中納言バージョン 2.7.2, バージョン 2026.3) <https://www2.ninjal.ac.jp/cojads/index.html> (2026 年 8 月 7 日確認)
- 国立国会図書館デジタルコレクション <https://dl.ndl.go.jp/> (2026 年 8 月 7・8 日確認)
- NDL Ngram Viewer <https://lab.ndl.go.jp/ngramviewer/>
- ジャパンナレッジ Lib <https://japanknowledge.com/library/aboutlib.html>