

『現代日本語書き言葉均衡コーパス』「図書館書籍」の生年代別分布は何を表しているのか —「デナイ」「デハナイ」「ジャナイ」の使用割合から見た一考察—

森 秀明 (東北大学大学院文学研究科)

What Does the Birth Year Distribution in The Library Sub-Corpus in the BCCWJ Represent; A Study Based on the Percentage of “denai”, “dewanai” and “janai”

Hideaki Mori (Graduate School of Arts and Letters, Tohoku University)

1. はじめに

『日本語話し言葉コーパス』(以下 CSJ と呼ぶ) や『現代日本語書き言葉均衡コーパス』(以下 BCCWJ と呼ぶ) には、生年代情報が付与されているサンプルがある。分析者がこれを利用して、調査対象の生年代別分布を調べることも多いだろう。しかし、そもそもこれらのコーパスで得られた生年代別分布は何を意味しているのだろうか。また、その分布を無批判に分析に利用しても良いものだろうか。ここでは前半で二つの先行研究を取り上げ、生年代別分布の捉え方や分析の問題点を考察する。後半では具体的な言語形式を取り上げ、BCCWJ の「図書館書籍サブコーパス」(以下「図書館書籍」と呼ぶ) を利用して生年代別分析を行い、実年代データと比較しながらその分布が何を意味するのかを考える。

2. 先行研究

2. 1 佐野(2012)における生年代別分布の捉え方と分析の問題点

佐野(2012)では陳述副詞「全然」の用法において「全然大丈夫、全然いい」などの「革新的」な用法が近年増加しているかどうかを確かめるため、CSJ を使用した生年代別分析が行われている。この分析に当たって佐野(2012)では「話者の生年における差を時間の流れに見立てて考える(見かけ時間)」(p. 36)として「見かけ時間」という捉え方が提示されている。

見かけ時間とは、時間軸上のある一点で言語データの収集を行い、その分析結果に見られる年齢差や世代差から時系列的な変化を推定する考え方である。この考え方の基盤には、人間が言語形成期(10歳前後)に獲得した言語は終生保持されとする臨界期記憶仮説がある。実時間で何度も調査を行うには膨大な時間や費用がかかるほか調査対象者の確保など様々な問題も多いため、社会言語学では多用されている分析手法である。

CSJ の調査年は非公開であるが、佐野(2012)はこれを時間軸上のある一点で言語データの収集を行ったものとみなして、見かけ時間で捉える立場を取る。その上で「全然」の使用法の「経年変化」をまとめたものとして、図1を示している。

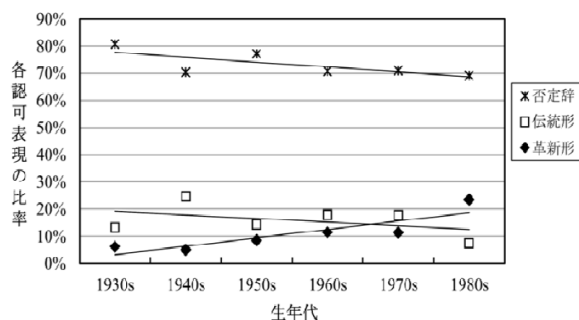


図1 「全然」の呼応表現別経年変化
(佐野, 2012, p. 36 より転記)

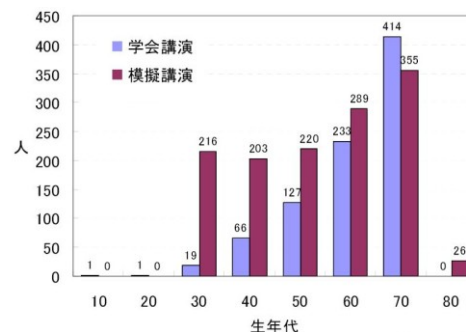


図2 学会講演と模擬講演話者の生年による分布
(前川, 2006, p. 6 より転記)

図1の「否定辞」とは「全然～ない」などの用例、「伝統形」とは「全然～違う」などの用例、「革新形」とは「全然～いい」などの用例を指す。

この分析を検討するために、簡単にCSJの性格を見ておく。CSJとは、国立国語研究所等が作成した、日本語の自発発話の音声を格納した大規模コーパスである。録音音声ファイルのほか、文字化して形態論情報が付与されたXML文書も提供されているため、文法や意味の分析にも利用できる。語数はおおよそ750万語で、その内訳は学会講演が328万語、模擬講演が360万語、その他が62万語となっている。学会講演とは理工学を中心とした研究発表のライブ録音で、講演者は男性の大学院生が多く、あらたまり度の高い発話が多い。模擬講演は人材派遣会社から派遣された一般話者が、「人生で一番うれしかったこと」等について行った10～15分程度のスピーチで、学会講演よりもくだけた発話になっている。

このようなCSJの性格からすると、佐野(2012)の言う「革新形」は学会講演より、模擬講演で多く用いられていることがあらかじめ予想される。それを考慮に入れた上で、前川(2006)に記載されているCSJの講演種類別の講演者の人数を見てみよう。図2を見ると、学会講演と模擬講演で人数に著しい差があることが分かる。しかも1980sの人数はわずか26人しかいない。さらにその全てが模擬講演を行っている。これを図1で確認すると、やはり1980sのみ「革新形」の割合が高い。この年代を図1から除いた場合、「革新形」の推移はほぼ並行になっているようにも見える。これで果たして「近年多く見られるようになった革新形が確かに徐々に増加してきていることが実証された。」(p. 37)と言えるであろうか。

このような分析結果となったのは、もともとのコーパスが持っているデータの偏りを検討せず、無批判に分析に使用したためと考えられる。もし、各年代ごとに比較的人数が安定している模擬講演だけで分析を行ったら、もっとはっきりした傾向が見えた可能性がある。ただし、分析対象を模擬講演だけに絞った場合、今度は話し言葉の一部の位相でしか時系列的な変化を見ていないことになる。

見かけ時間とは、年齢差や世代差から時系列的な変化を推定することであった。しかし、何をどれぐらいの精度で推定するかは、それぞれの分析目的によって異なるだろう。一口に話し言葉と言っても、そこには様々な姿がある。一人の人間が話す言葉でも、聞き手の年齢や地位、あるいはプライベートな場かフォーマルな場かなどによって使用される言葉は変化する。これら様々な社会的要因によって特徴づけられた言語使用の現象を位相と呼ぶなら、できるだけ多くの位相を対象にして現代日本語の平均的な姿を推定しようとするか、言語変化が現れやすい一部の位相を対象にして変化の傾向を推定しようとするかでは、結果が当然異なる。佐野(2012)ではCSJの全データを分析対象としているから、フォーマルな話し言葉も含めた現代日本語でどのような変化が起きているかを調査する目的があったと思われる。もしそうであれば、学会講演と模擬講演のデータ数のばらつきを補正した上で分析すれば、より「推定したい現代日本語」に近似した結果が得られた可能性がある。

一方、分析の精度についてはどうだろう。CSJでは調査年代が非公開になっているが、これを仮に2000年に行われたものだと仮定すると、生年代1970sの世代は調査当時30代、1930sは70代になる。実時間で2000年の「現代日本語」の姿とは、おおよそこれら30代～70代の平均的な言葉遣いだといえよう。この平均が仮に40代のものと同じだとしたとき、70代はおおよそ30年前、30代はおおよそ10年後の平均的な言葉遣いを表していると推定される。もし2010年にCSJと同様の調査を行い、その平均が30代の言葉遣いと一致すれば、この見かけ時間による推定は精度が高かったことになる。

しかし、このような推定の精度が必ずしも高くないことは日比谷潤子編著(2012)などの入門書でも繰り返し指摘されている。見かけ時間を実時間と見なす上での問題点を横山(2011)で示された枠組みを利用して確認しよう。図3は横山(2011)で図示されている言語運用能力の生涯習得モデルである。言語運用能力はいつ生まれたかに最も影響される。これは臨界期記憶仮説をもとにした要因で「世代効果」と言える。さらに年齢を重ねた分の「加齢効果」と、調査した年代の言葉遣いの影響である「時代効果」が加わる。この二つの効果は調査した年によって決定されるため「調査年効果」とまとめることができる。先の仮定

に沿って生年代 1930s の例で考えれば世代効果とは $1930s + 10 \text{ 歳} = 1940s$ 頃に言語形成期を迎えたことによる言葉遣いの影響、加齢効果とは平均の 40 代より 30 年分年齢を重ねたことによる知識の増加や 70 代にふさわしい言葉遣いを選択することなどによる影響、時代効果とは 2000 年当時の平均的な言い回しや、流行語などによる影響ということになる。

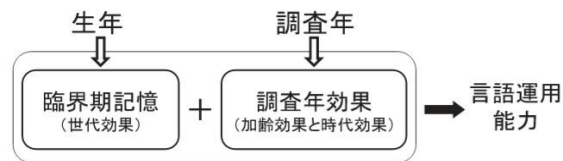


図3 生涯習得モデルの枠組み（横山, 2011, p. 29 より転記）

佐野(2012)では、調査年効果による影響が考慮されていない。CSJ の生年代別比較では時代効果は同一と仮定して無視できるとしても、加齢効果が使用割合に影響している可能性がある。やはり「実証」というのは難しく、おおまかな推定を示した分析と言えよう。

以上、佐野(2012)から生年代別分布を見かけ時間で捉える考え方があること、その分析に当たっては調査年効果の影響やコーパスに潜在しているデータの偏りを考慮する必要があることを述べた。

2. 2 前川(2012)における生年代別分布の捉え方と分析の問題点

前川(2012)では、「大きいです。」「小さいです。」のように、形容詞に助動詞の「です」が直接後続して文末を形成しているタイプの述語の成立要因を、BCCWJ を用いて検討している。この 4.5 節に「A（形容詞）＋です。」述語の社会言語学的動向を知るために、「使用者の年齢を検討する」として、生年代別分布の分析が行われている。これは生年代別分布を見かけ時間（いわば横の変化）で捉える立場を取らず、年齢による差（いわば縦の変化）で捉える考え方だと言えよう

前川(2012)では、「図書館書籍」と「ベストセラー」を用いて分析が行われているため、はじめにこれらの性格を確認しておく。「図書館書籍」は BCCWJ の中核をなすサブコーパスの一つで、語数はおよそ 3,000 万語である。東京都内の公立図書館に所蔵されている書籍のうち、1986 年から 2005 年の 20 年間に出版されたものを対象とし、ランダムサンプリングによってデータが集積された。書き言葉（書籍）が社会に流通している実態を公立図書館の所蔵状況によって近似的に捉えたもので、「流通実態サブコーパス」とも呼ばれる。出版年とジャンルに偏りが出ないように構成比を割り当てて 10,551 個のサンプルを選び、そこからさらにランダムに範囲を指定して各サンプルごとに約 1,000 文字を抽出している。このようにサンプルごとの文字数まで均等に抽出したデータを「固定長サンプル（約 670 万語）」と呼ぶ。この他に章や節などのまとまりのある範囲で最大 1 万字のデータを抽出したものを「可変長サンプル（約 2,889 万語）」という。この可変長サンプルはサンプルごとに語数のばらつきがあるため統計分析に使用する場合は注意が必要となる。これら固定長サンプルと可変長サンプルはその一部が重なっているものが多いため、この二つを足して重なりを除いた語数が約 3,000 万語となる（このサンプルを可変長、固定長に対して「両方」と呼ぶことにする）。

一方「ベストセラー」は、無作為抽出を施していない「特定目的サブコーパス」の一部を構成する、いわばサブコーパスのサブコーパスとでも言うべきものである。1976 年から 2005 年までの 30 年間に出版された書籍のうち、『出版年鑑』などで年間 20 位に入った書籍 951 冊のうち、「作業上の理由」から 695 冊を選び、可変長のみを抽出した（以上、詳しくは『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版参照のこと）。

まず、生年代を年齢と捉える考え方から検討する。「図書館書籍」のサンプル数は、10,551

個である。ただし、サンプルには生年代が不明のものや共著・共訳による書籍がある。共著・共訳ではすべての執筆者の生年代が一致する場合（以後これを「同年代共著」と呼ぶ）のみが生年代分布の分析に使用できる。ただし BCCWJ の Web 検索サイトの『中納言』で公開されている資料（BCCWJ_WC_LUW_v10.xlsx）には、翻訳者・共著者の区別ができる情報がないため、ここでは単著で生年代が判明しているサンプルのみを扱う（表 1）。これにより生年代が判明した 7533 個のサンプルを生年代別（年齢別）に示したのが図 4 である。

表 1 「図書館書籍」生年代の
判明・不明サンプル数

判明	不明	翻訳・共著	合計
7533	1164	1854	10551
71.40	11.03	17.57	100.00

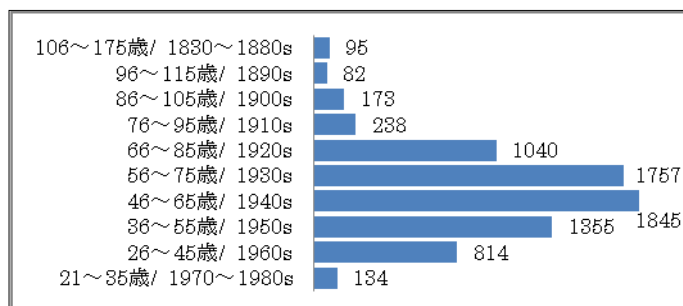


図 4 「図書館書籍」の生年代別サンプル数（翻訳・共著を除く）

縦軸にしている「年齢/ 生年代」を検討してみる。「図書館書籍」は 20 年間に渡って出版された書籍からサンプリングしている。このため一つの生年代における年齢の幅は、20 歳あることにある。理論的には 1970~1980s は 6 歳~35 歳となるが、1970s は 1993 年以後、1980s は 2001 年以後のものしかない。これはいわば「観察打ち切り」のデータでサンプリング対象が 2005 年で区切られているため、20 歳分のデータが集積できなかったことによる。

ここから縦軸を順に見ていくと、1910s 以上は実際に執筆した年齢とは見なしにくい。特に最上段の 1830~1880s では人間が生存して執筆できる年齢とは考えられない。これと横軸のサンプル数を比べると、1910s から極端に少なくなっている。個別に確認しないと確実ではないが、1910s 以前の生年代は、生前に執筆した書籍が再出版されている可能性が高い。そのように見なしたときこれらの生年代は全生涯のかなりの部分に渡る年齢となる。

以上のように検討すると生年代を年齢で捉えた場合、およそ 20 年幅の年齢と全生涯のかなりの部分に渡る年齢とが混在することになる。年齢という一つの指標に 2 つの基準が混在するため、生年代を年齢で捉える考え方には問題があると言えるだろう。

次に前川(2012)の分析内容を検討する。前川(2012)の分析結果を転記したものが図 5 である。図 5 では横軸が生年代となっているが、前川(2012)ではこれをあくまで年齢で見ている。

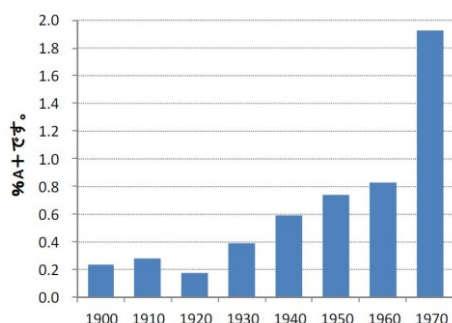


図 5 書き手の生年代と「%A+です。」の関係
(前川, 2010, p. 220 より転記。縦軸は%)

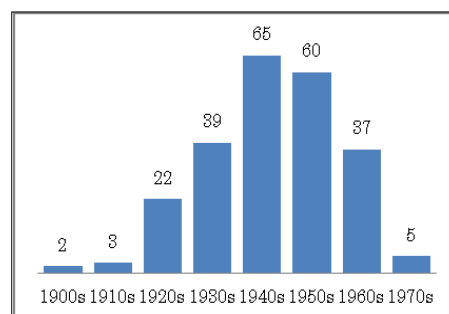


図 6 「図書館書籍」+「ベストセラー」における「A です。」の頻度（縦軸は用例頻度）

「%A+です。」とは、「A。」の頻度と、「A+です。」の頻度合計に対する「A+です。」の割合を指す。この分析から前川(2012)では「書き手の年齢が低下するにつれて「%A+です。」

述語の使用率が上昇する」ことを指摘した上で、「特に1960年代から1970年代にかけて著しく上昇していることは注目に値する。」と述べている。

図6は図5のもとになった「A+です。」の頻度を再調査した結果である。調査には『中納言』を使用し、検索対象を「図書館書籍」+「ベストセラー」に限定した上で、キー: 形容詞、後方共起1: 書字出現形「です」、後方共起2: 書字出現形「。」、サンプルは「両方」で検索した。なお検索データの分析には田野村忠温氏が開発し、無償でネット公開されているsortKWICを使用した(3節の分析も同じ)。

図6を見ると、両端の生年代で著しく頻度が小さい。特に前川(2012)で「注目に値する。」と述べられた1970sの頻度はわずか5例しかなく、これで1%の有意差を評価するのは統計的に困難だと思われる。そもそも両端の頻度が極端に少なくなるのは、分析対象としたコーパスが持っている分布の偏り(図4)のためである。「図書館書籍」では、出版年やジャンルには厳格な構成比の割り当てがなされているが、生年代においてこの点は考慮されていない。生年代別分布の分析に当たっては、特にコーパスに潜在している分布の偏りに留意する必要がある。さらに1970s以後は、「観察打ち切り」によって十分なデータが集積されていない。この年代に特異な振る舞いが観察された場合は、慎重な判断が必要であろう。

以上、前川(2012)の検討から「図書館書籍」の生年代を年齢で捉えるのは問題があること、生年代別分布の分析に当たっては、コーパスに潜在しているデータ分布の偏りや観察打ち切りデータの扱いに留意すべきことを述べた。

2. 3 「図書館書籍」を見かけ時間で捉える場合の問題点

佐野(2012)と前川(2012)から生年代分布の捉え方には見かけ時間と年齢という二つの捉え方が分かった。しかし「図書館書籍」の場合、これを年齢で捉えると一つの指標に二つの基準が存在することになり問題があった。それでは「図書館書籍」を見かけ時間で捉えることは妥当なのだろうか。CSJはある一点と見なしうる調査年代に実際に発話されたデータを集積したコーパスだが、「図書館書籍」は20年に渡って出版された書籍を集積している。しかも1910s以前は生前に出版された書籍が再出版されている可能性がある¹。

この問題を図3の生涯習得モデルの枠組みから考えると、世代効果は一義的に決まるが、加齢効果は1920s～1960s頃までは10年ごとの重なりを持つ20歳幅の効果、1910s以前は全生涯のかなりの部分における各加齢効果の合計となる。図3の調査年を出版年とした場合、そもそも1910s以前は、出版年によって加齢効果が決まるわけではない。これは時代効果も同じで1920s以後は出版年である1986年～2005年の時代効果の影響を受けていると考えられるが、1910s以前は執筆者の生年-没年で時代効果が異なる。

このように考えると生年代別分布を見かけ時間で捉えた場合、理論的には解決が難しい問題が存在する。理論的にはこれ以上踏み込めないため、次節ではケーススタディを行って生年代別分布と実時間データの分布を比較し、生年代別分布を見かけ時間で捉えた場合、実際にどのような問題が生じるのかを確かめてみる。

3. ケーススタディ

3. 1 用例の抽出

検索対象はBCCWJの「図書館書籍」、調査対象の言語形式には名詞述語とナ形容詞述語の否定形に使用される「デナイ」「デハナイ」「ジャナイ」(以後「コピュラ否定形」と呼ぶ)を使用し、その生年代別分布を観察する。コピュラ否定形を使用するのは、生年代によって使い分けがあるものの、その合計数はあらゆる生年代で一定であると考えられるからである。一般的に考えて、特定の年代で動詞述語やイ形容詞述語だけが使われやすかったり、肯定形に比べて否定形が少なく使用されるなどの変動は考えにくいだろう。

¹1910s以前に限らず、書籍にはそもそも執筆年と出版年が必ずしも一致しないという問題がある。丸山(2012)によれば、書籍の大部分は初出から3年以内に出版されているため、ここではその問題には触れない。

ただし、名詞述語とナ形容詞述語の否定形には丁寧体において「デハアリマセン」等も使用される。「図書館書籍」では、これらがコンピュータ否定形の1割程度使用されているが、これは今回の分析の対象とはしない。ただし、同じ丁寧形でも「デハナイデス」等の形式は含まれる。

用例の抽出には『中納言』を使用して検索対象を「図書館書籍」に限定し「デナイ」の場合、キー: 未指定、後方共起 1: 書字出現形「で」、後方共起 2: 語彙素「無い」+品詞「形容詞」で検索した。サンプルは「両方」、単位は「短単位」を選択した。このように抽出したデータから単著で生年代情報が得られているものに加え、単独の訳者による翻訳、同年代共著を含めて形式・年代別に示したのが表2である。以後の分析はこれをもとに行う。

表2 「図書館書籍」におけるコンピュータ否定形の形式・生年代別頻度

形式	1880s ~	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s ~	小計	不明	合計
デナイ	150	129	307	359	1270	1763	1854	1211	753	76	(7872)	1460	9332
%	30.80	22.09	23.85	21.23	19.28	14.76	15.36	12.11	12.19	8.47	(15.22)	17.34	15.51
デハナイ	287	333	787	1052	4370	7552	7601	6015	3362	458	(31817)	4654	36471
%	58.93	57.02	61.15	62.21	66.34	63.21	62.95	60.16	54.42	51.06	(61.50)	55.28	60.63
ジャナイ	50	122	193	280	947	2633	2619	2773	2063	363	(12042)	2305	14348
%	10.27	20.89	15.00	16.56	14.38	22.04	21.69	27.73	33.39	40.47	(23.28)	27.38	23.85
合計	487	584	1287	1691	6587	11948	12074	9999	6178	897	(51732)	8491	60151
%	0.81	0.97	2.14	2.81	10.95	19.86	20.07	16.62	10.27	1.49		14.00	100.00

* 「~1880s」にはそれ以前の年代、「1970s~」にはそれ以後の年代を含めた。

* 生年代情報がないもの、共著・共訳で執筆者の年代が異なるものは不明とした。

3. 2 実時間データとの比較

実時間データと比較する前に、「図書館書籍」の場合、何をもって実時間データとみなすのかを定義しておく必要がある。一般に実時間と言った場合、書籍が執筆された年と考えやすい。しかし「図書館書籍」は、そもそも流通実態を反映するために設計されたコーパスである。古い時代に執筆出版された書籍が再出版された場合、そこで使用されている言語も流通実態としては出版年に流通している言語の一部を構成している。そこでここでは「図書館書籍」における実時間の使用言語を「ある年代に出版され図書館に所蔵された書籍の総体に使用されている言語、あるいはそれらの平均」と定義する。なお、古い年代においては『太陽コーパス』を実時間データとみなして使用する。『太陽コーパス』は、2005年に国立国語研究所が刊行した近代語研究のためのコーパスである。明治末期から大正末期にかけて、当時最もよく読まれ、多彩なジャンルに渡る記事を掲載していた雑誌『太陽』を、1895,1901,1909,1917,1925年の各年ごとにコーパス化したもので、約1400万字が収録されている。『太陽』は一つの雑誌に過ぎないが、これまでの先行研究において当時の書き言葉の使用実態を反映していると評価されているため、これをここでは実時間データとみなすことにする。

実時間データの『太陽コーパス』と「図書館書籍」の出版年別にコンピュータ否定形の頻度を調査した結果が表3、表4である。表4の「図書館書籍」出版年は1986~1989年の4年分を1988年、1990~1999年の10年分を1994年、2000~2005年の6年分を2002年と表示した。これらを使って、表2の「図書館書籍」の生年代分布との比較を行う。

表2のデータと表3、表4のデータを重ね合わせてグラフを描いたものが図7である。ただし、図7において実時間と見かけ時間の時間幅は同じではない。また、グラフを重ねるに当たっては最も割合が近似すると思われる年代に重ねたが、この重なりが直ちに同年代を意味するものではない。あくまで生年代別分布を何の補正もせず、単純に見かけ時間と

見なした場合、実時間とどのような関係になるかを比較するためのグラフである。

表3 『太陽コーパス』の出版年別分布

形式	1895	1901	1909	1917	1925	合計
デナイ	162	394	986	1047	843	3432
%	38.39	46.08	45.04	40.52	32.98	39.88
デハナイ	197	316	947	1152	1427	4039
%	46.68	36.96	43.26	44.58	55.83	46.93
ジャナイ	63	145	256	385	286	1135
%	14.93	16.96	11.69	14.90	11.19	13.19
合計	422	855	2189	2584	2556	8606

表4 「図書館書籍」の出版年別分布

形式	1988	1994	2002	合計
デナイ	1222	5046	3064	9332
%	17.53	15.64	14.65	15.51
デハナイ	4095	19574	12802	36471
%	58.75	60.65	61.23	60.63
ジャナイ	1653	7653	5042	14348
%	23.72	23.71	24.12	23.85
合計	6970	32273	20908	60151

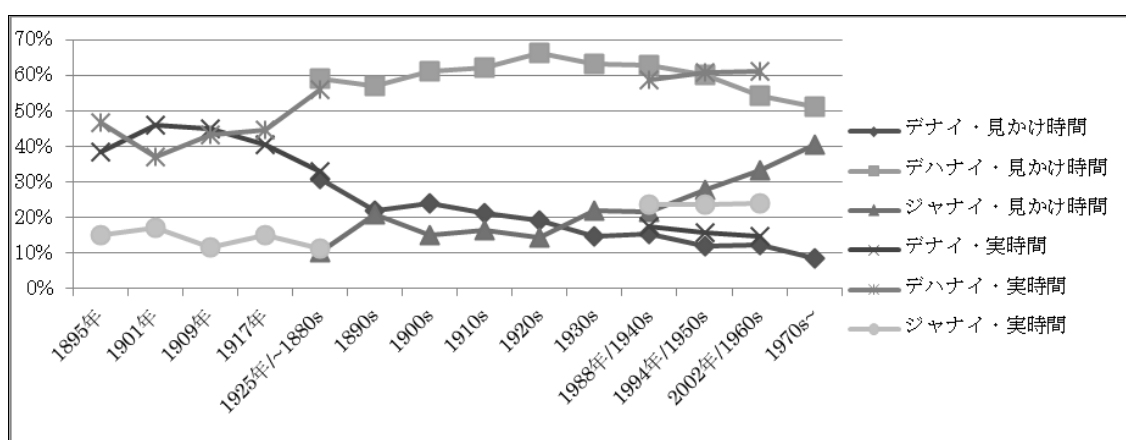


図7 実時間『太陽コーパス』、見かけ時間「図書館書籍」生年代、実時間「図書館書籍」出版年の比較

初めに実時間から検討する。実時間では1895~1917年頃まで、デナイとデハナイが同程度の割合で使用されている(1895年と1901年でグラフが変動するが、この頻度を表3で確認するとこの区間のみ他の区間の1/6~1/3程度しかない。このためこの変動はデータ不足による可能性が高いと思われる)。それが1925年になると大きく使用傾向が別れ、デナイよりデハナイを使用する割合が増えている。この時期は言文一致運動の成長期で、特に1922年には全国誌の新聞でも全紙面で言文一致を採用するに至っている。この変化はそのような時代の流れを反映したものと思われるが、これについては稿を改めて述べる。

一方、1988~2002年(正確には1986~2005年)の区間では、ジャナイは24%程度でほぼ一定で、デハナイが6割弱から6割強へ微増、デナイが17%半ばから14%半ばへ微減している。わずかな変化は見られるが、ほぼこの区間のコピュラ否定形の使用割合は一定だと言えるだろう。これらは実時間のデータをもとにした分析であり、流通実態としての書き言葉の平均的な姿を、ある程度正確に反映していると考えられる。

次に見かけ時間を検討する。~1880s~1940sまでは、生年代によって多少のこぼこがあるもののほぼ一定の傾向を示している。デハナイは6割程度であり変化がない。デナイは3割程度から徐々に半減していく。ジャナイは1890sで特異に増加した後はあまり増加せず、1920sで急に増加した後、1930sから1940sの区間ではまた水平になる。全体的な傾向ではデハナイはほぼ横ばい、デナイは30%から15%へ半減、ジャナイは10%から20%強への増加となっている。~1880sから1910sまでの区間はデータ数が少ないものの、1890sの特異な変動を除けば、実年代の1917年から1925年までの傾向をそのまま自然に推移させた姿のように見える。また3.2節の考察によれば、1910sと1920sには加齢効果と時代効果に違いがあるはずである。しかし図7ではこの区間で大きな変化があるようには見えない。

一方、1950s以降はデハナイとジャナイに急激な変化が見られる。デナイはそれまでの傾向通り徐々に割合を下げていくが、デハナイは1940sの6割強から1970sの5割強へおよそ10%減少、ジャナイは同様に21%強から40%強へ増加している。これを実時間と比較すると、大きく異なる値となっているのが分かる。表2で頻度を確かめても、1950sと1960sには十分な頻度が確認されることから、この区間における実時間と見かけ時間の乖離は、データ不足が原因ではないと考えられる。

以上、実時間と見かけ時間を比較することにより、見かけ時間の1940s以前は、1890sの特異な変動を除けば、およそ実時間の傾向のまま推移しているように見えるが、1950s以降は実時間と大きく乖離していく傾向があることを観察した。

3.3 ジャンル効果

前節では見かけの時間の1950s～1970sの区間が、実時間と乖離していく傾向を観察した。しかし、なぜこのような傾向が観察されるのだろうか。言語変化は若い世代ほど起きやすいと言われている。その加齢効果が強く影響しているのであろうか。それにしてもあまりにも急激な変動に思われる。そこで加齢効果以外の要因を検討するため、まずジャンルごとにコピュラ否定形の使用割合に変化があるかどうかを調査した。

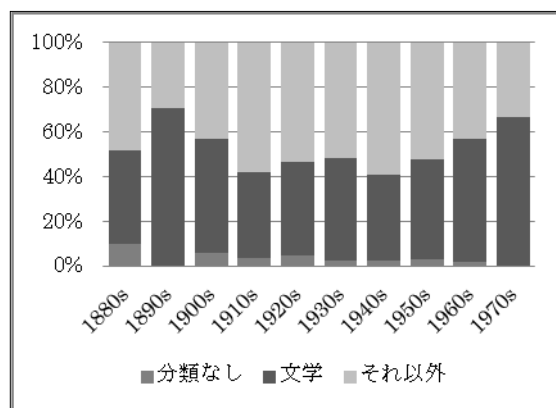
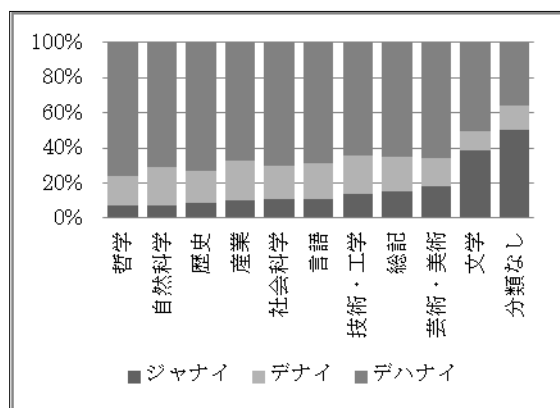


図8 「図書館書籍」ジャンル別コピュラ否定形割合 図9 「図書館書籍」生年代別ジャンル構成比率

図8を見ると文学と分類なしで他のジャンルの数倍ほどジャナイが使用されていることが分かる。分類なしには明らかに小説と考えられる書籍が多く含まれているから、基本的に会話の描写が多い書籍でジャナイが使用されやすいと考えて良いだろう。

次に図9で生年代別のジャンル構成を見ると、1950sから次第に「文学+分類なし」が増加していることが分かる。図7の見かけ時間において1950s～1970sの年代でジャナイの比率が増加していたのはこの影響による要因が大きいと考えられる。ただし、図9ではこの年代以外にも偏りが見られる。1880s～1900sでは総じて「文学+分類なし」の割合が高く、特に1890sの偏りが著しい。これを図7で確認すると、この年代だけ特異にジャナイの比率が高くなっている。これを補正すればこの区間はもっとなめらかに推移すると考えられる。

「図書館書籍」で古い年代の文学比率が高くなるのは、時代を超えて読み継がれる書籍には文学が多いからだだろう。一方、新しい年代で高くなるのは若い世代が書籍を出版できるとしたら、専門書などより文学の可能性が高いからであろう。これら「図書館書籍」に潜在する要因で、分布が偏っていたのである。

以上から、コーパスの生年代別分布には、ジャンルによる対象言語の使われやすさ（ジャンル別使用傾向）と、生年代別のジャンルの偏り（生年代別ジャンル偏向）が影響していることが分かった。さらに「図書館書籍」でジャンルごとの可変長語数を比較すると、最も少ない「言語」に対して最も多い「文学」では約1.7倍になっている。このため検索

オプションを固定長以外にして検索すると、この影響も受ける。これを（ジャンル別可変長偏向）と呼ぶことにする。これらは連動して働き、分布に影響を及ぼす。この3つを一括してジャンル効果と呼ぶと、コーパスの生年代分布はジャンル効果によって影響されていると考えられる。

3. 4 調査年効果

2. 3節で考察したように1910s以前の加齢効果と時代効果は「図書館書籍」の出版年とは無関係であると考えられる。これらの調査年効果を調べるには個別のサンプルを一つ一つ確認していくしかない。BCCWJのDVD版には書誌情報が収録されているが、丸山(2012)によれば初出・初刊の情報はおよそ25%ほどしか取得できないため、この区間の調査年効果を正確に観察することは難しい。また、1970s~は1993年以後のサンプルしかない。よってここでは1920s~1960sのみを対象に調査を行う。

調査は表2のデータを5年間の出版年ごとに区切り、コピュラ述語否定形の割合をグラフに描いて観察する。本来1年ごとに比較を行うべきだが、表2のデータを20分割すると頻度が少なくなる上に、グラフの線が重なりすぎて分かりにくくなる。このため5年分ずつの比較とした。図10~12の1986-等の表示は1986-1990年等の5年間であることを表す。

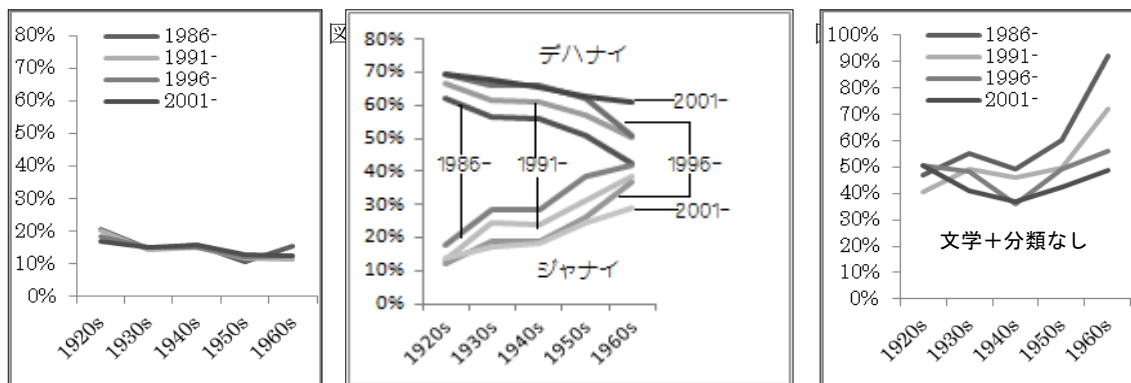


図10は、デナイだけを取り出したグラフである。図10ではグラフがほぼ一直線にまとまることから、デナイには調査年効果がほぼ観察されない。デナイの使用割合が緩やかに減少していく要因は、生年代による世代効果によるものと考えられる。図11はデハナイとジャナイの割合のみを示したグラフである。これを見ると互いがほぼ相補的に推移していることが観察される。1960sを除くと、1986-1990年、1991-1995年、1996-2000年の2つの間隔は全ての生年代でほぼ等しくなっている。また1996-2000年と2001-2005年は全ての年代で重なっている。これは1920s~1950sの生年代にほぼ同一の効果が及んでいることを示唆している。調査年効果には、時代効果と加齢効果があるが、加齢効果は各生年代で異なっていると考えられるため、このように均一に働く効果は時代効果である可能性が高い。1986-1990年、1991-1995年の各5年間は、ジャナイに変わってデハナイが使われやすくなる傾向が見られるが、1996-2000年の5年間でこの傾向が見えなくなる。

このような傾向にジャンル効果が関わっているかどうかを確認するために、5年区切りの出版年別に文学+分類なしのジャンル割合を調査したのが図12である。これと図11を比較すると、図11にジャンル効果が一定程度及んでいるように思われるが、図11の均等な傾向は図12のジャンル割合では説明できない。このため図11の傾向はジャンル効果以外に時代効果が働いていることを示唆していると考えられる。

なお、1960sは1996-2000年の5年間でジャナイの割合が特異に高くなる。これはこの生年代のこの区間だけに観察される現象である。このためこの区間の用例を個別に確認した。すると、文学や分類なし以外のジャンルでも会話が多用されている傾向が見られた。図12

ではこの区間の文学+分類なしのジャンル割合が 56%しかなく、1986-1995 年に比べて低いように思われる。しかし 1 年ごとに調査すると、例えば 1996 年では文学+分類なしのジャンル割合が 48%しかないのに、ジャンナイの割合は 49%になっているなどの現象が観察された。1996 年の場合、社会科学に分類されている『中国てなもんや商社』、歴史に分類されている『超貧乏旅』、美術・芸術に分類されている『リチャードレスター』などで会話が多用され、ジャンナイが多く出現する。1960s で 1996-2000 年の区間が特異なふるまいを見せる原因には、ジャンナイが多用されるサンプルが集積されたことが考えられる。

以上、調査年効果の調査から 1920s~1960s の生年代では、時代効果が観察され、1986-1995 年の 10 年間ではデハナイが一定の割合で増加する傾向が見られるが、1996 年からはその傾向が見えなくなることを述べた。

4. まとめ

先行研究では、生年代別分布を見かけ時間で捉える立場と年齢で捉える立場があった。「図書館書籍」では、およそ 1910s を境に年齢の基準が変わるため、生年代分布を年齢で捉えることは難しい。これを見かけ時間で捉えた場合にも 1910s を境に調査年効果に変化すると考えられる。本稿ではケーススタディを行って「図書館書籍」の生年代分布を見かけ時間で捉えた場合の問題点を考察した。

その結果、生年代分布に作用する因子として、ジャンル効果と時代効果が観察された。ジャンル効果とは、コーパスが潜在的に持っている分布の偏りのため、特定の生年代にあるジャンルが偏り、そのジャンルに現れやすい言語の頻度が高くなる効果のことである。言語の変化は書き言葉より先に会話に現れやすいため、会話の描写が多い文学の比率が高まると、その年代であたかも言語変化が進行しているかのように見える。

コピュラ否定形における「図書館書籍」の生年代別分布は、ジャンル効果などの影響により、無意味な内容を表している。コーパスに潜在する分布の偏りを補正することで、生年代別分布ははじめて意味を持つと言えるだろう。

参考文献

- 佐野真一郎(2012)「『日本語話し言葉コーパス』を用いた「全然」の変化の詳細化」『第 1 回コーパス日本語学ワークショップ予稿集』 pp. 33-42.
- 日比谷潤子編著(2012)『はじめて学ぶ社会言語学—言葉のバリエーションを考える 14 章—』ミネルヴァ書房
- 前川喜久雄(2006)「第 1 章 概説」『日本語話し言葉コーパスの構築法』 pp. 1-18.
- 横山詔一(2011)「言語変化は経年調査データから予測可能か?」『国研プロジェクトプレビュー』 No.6 pp. 27-37.
- 前川喜久雄(2012)「「形容詞+です」述語の生起要因についての準備的考察」『第 1 回コーパス日本語学ワークショップ予稿集』 pp. 211-220.
- 丸山岳彦(2012)「大規模コーパスの利用とメタデータの役割」『第 1 回コーパス日本語学ワークショップ予稿集』 pp. 203-210.

関連 URL

- 『中納言』BCCWJ 短単位語数 https://maro.ninjal.ac.jp/wiki/index.php?BCCWJ_短単位語数
- 『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版
http://www.ninjal.ac.jp/corpus_center/bccwj/doc.html
- 『第 1 回コーパス日本語学ワークショップ予稿集』
http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no1_papers/JCLWorkshop2012_web.pdf