国立国語研究所学術情報リポジトリ

読点が接続詞の直後に打たれる要因について ： 一般化線形モデルを用いた予測モデルの構築

著者	岩崎 拓也
雑誌名	言語資源活用ワークショップ発表論文集
巻	2
ページ	56-63
発行年	2017
URL	http://doi.org/10.15084/00001506

読点が接続詞の直後に打たれる要因について：一般化線形モデルを用いた予測モデルの構築

岩崎 拓也（一橋大学大学院言語社会研究科）

Factors Involved in Placing the Comma Immediately Following a Conjunction: An Analysis Using Generalized Linear Models

Takuya Iwasaki (Hitotsubashi University Graduate School of Language and Society)

要旨

本稿では，接続詞の直後に読点を打つ場合の要因を探るためにコーパスを用いた計量的な調査を試みた。コーパスには，『現代日本語書き言葉均衡コーパス』のコアデータを使用し，検索には『中納言』を用いた。分析は，一般化線形モデルを用いたロジスティック回帰分析から，接続詞の直後に読点が打たれる要因について検討した。その結果，接続詞の直後に打たれる読点は，接続詞の連接類型だけでなく，接続詞の前後の文字種や品詞，接続詞の語種，一文の長さや接続詞自体の文字数といった変数が影響を与えていることが確認された。その中でも，一文が長いほど接続詞の直後に読点が打たれやすいこと，接続詞の文字数が多いほど読点が打たれにくいこと，接続詞の連接類型が対比型と同列型の時に読点が打たれにくいことが明らかになった。

１．はじめに

日本語の句読点には，正書法というものが確立していないと言われているが，実際には，文部省（1963）「付録 句切り符号の使ひ方〔句読法〕(案)」という規則が存在している[1]。ところが，現実の使用場面においては，句読点の使用規則を遵守しているとは言いがたい。文部省（1963）では，以下の例文(1)と(2)のように，接続詞の後に読点が打たれている場合と打たれていない場合の例がそれぞれ記されており，どちらがより使用されるのかといった言及は見られない。

(1) しかし私は、
(2) しかし、私は……

（文部省（1963））

岩崎（2016）では，日本語母語話者と上級日本語学習者が日本語で書いた作文の句読点の多寡を比較し，それぞれの差異を明らかにしている。また，岩崎（2017 予定 a）では，品詞別による読点の打たれやすさについて明らかにした上で，最も偏りのあった助詞の中から係助詞と格助詞について，係り受けの関係から分析を行っている。さらに，岩崎（2017 予定 b）では，接続助詞と読点の関係について，従属節の独立度の関係から考察を行っている。

他の先行研究を見てみると，金・樺島・村上（1993）が，文学作品と科学技術論文をそれぞれ対象とし，読点の打ち方により書き手の個性が判断できるか，読点とその直前の一文字

1 「付録 句切り符号の使ひ方〔句読法〕(案)」は，当時の文部省教科書局調査課国語調査室が作成したもので，昭和 21 年（1946）に文部省から発行されたものである。文部省（1962）の付録は再録である。

を抽出してその文字の差異から分析を行っている。その結果，読点の前の文字や読点の前の文字の品詞，読点を打つ間隔に関する情報の有効性について作家ごとに特徴が見られると述べている。

小磯ほか（2008）では，BCCWJ を用いてジャンル間の文体差に関わる要因を判別分析している。その結果，名詞率や漢語率といった他の指標よりも，読点を打つ頻度（小磯ほか（2008）では「コンマ単位長」と表記）は，ジャンルごとに若干の差はあるものの，その差はあまり大きくなく，文章のジャンルよりも個人に帰属する可能性を示唆している。

石黒（2011）は，『日本語文章・文体・表現事典』の「14 句読点のルール」において，読点は「①意味，②長さ，③構造，④表記，⑤音調，⑥リズム，⑦強調などの要因が複雑に絡みあって打たれるものであるものであ」ると指摘している。また，誰もが読点を打つ箇所には，①意味，②長さ，③構造が関わっており，人により異なる箇所は，④表記，⑤音調，⑥リズムが関わっているとしている。また，一部の人が打つ読点として，⑦強調を挙げている。

岩崎（2017 予定 a）の分析では，品詞の直後における読点の打たれやすさについて，母語別と品詞間で比較を行った。その後，岩崎（2017 予定 a，2017 予定 b）では，有意差が確認された品詞のみを取り上げて分析対象として扱った。しかし，有意差が確認されなかった品詞は，他の品詞に比べて読点が打たれにくいだけであり，必ずしも読点が打たれないわけではない。上記の例文(1)や(2)のように，同じ接続詞であっても，その直後の読点の有無にはばらつきが存在しており，そこには何らかの変数が影響を与えていると考えられる。従来の句読点に関する研究では，意味や構造という視点から分析，考察が行われることが多く，他の変数についても取り上げて分析する必要がある。

そこで本研究では，従来の文章術的なアプローチではなく，実際の言語使用の実態に基づき，意味や構造以外のいくつかの変数を考慮した上で，接続詞の直後の読点使用について，どのような変数が読点の打ち方に影響を与える要因として存在しているのか，また，どのような影響を与えているのかを考察する。

２．研究対象

本研究では，『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）のコアデータを使用する。また，検索系として，コーパス検索アプリケーション『中納言』を用いた。キーに「品詞−大分類−接続詞」を選択，前後文脈の語数は 500 文字，検索対象は固定長と可変長の両方を選択，文脈中の区切り記号はなしに設定し，長単位検索を行い，ダウンロードした。以下に詳述するが，接続詞ではないと考えられるものと連接類型のうち，「順接型」と「転換型」の判断がつかないものを取り除いた 5547 例を今回の分析データとした。その後の処理と分析は R（ver.3.3.3）で行った。

３．指標の抽出とその傾向

まず，分析にあたり，分析対象となるデータから以下の 12 の指標を抽出した（表 1）。

接続詞直後の読点の有無は，BCCWJ コアデータから抽出したデータの後文脈から空白を削除した上で，一文字目を抽出し，確認を行った。その結果，読点なしは 2844 例（51.3%），読点ありは 2703 例（48.7%）であった。この指標を応答変数として以下では取り扱う。

次に，本研究で取り扱う説明変数について説明する。

読点を打つかどうかは文の長さに影響するという先行研究の指摘を受け，接続詞を含む文の一文の長さを抽出し，変数として設定した。

表 1 分析データから抽出した指標

・接続詞の直後の読点の有無（応答変数）
・接続詞が使用されている文の長さ
・文頭にあるか否か
・接続詞自体の文字数
・接続詞の類型
・接続詞の語彙素
・レジスター情報
・接続詞の語種
・接続詞の最初の一文字目とその前の単語の最後の一文字の文字種
・接続詞の最後から一文字目とその後の単語の最初の一文字の文字種
・接続詞の前に使用された品詞情報
・接続詞の後に使用された品詞情報

また，今回は接続詞の連接類型を説明変数として扱った。接続詞の連接類型判定については，市川（1978）の接続詞の連接類型の基準[2]を採用した。この基準をもとに別途，接続詞（語彙素）と連接類型を一致させた連接類型判定用データを作成し，この判定用データと分析対象データを結合させることで，分析対象データの接続詞の連接類型を判定した[3]。

語彙素とレジスター情報，語種については，BCCWJ に付与されている情報をそのまま流用する。ただし，語彙素については，副詞的に使用されていたもの（「さてさて（2 件）」）を削除した。また，語彙素の中でも表記のずれがあるもの（「しかーし（1 件）」「しっかし（1 件）」）などについては，目視にて基本的な語彙素と思われる表記のもの（先述した例の場合は「しかし」）に整理した[4]。文頭かどうかは，BCCWJ に付与されている情報を使用した。

---

[2] 市川（1978）では，文と文との論理的関係のことを「文の連接関係」と呼び，基本的な連接類型として8種を挙げている。本研究では，「文の連接関係の基本的な連接類型」のことを「連接類型」と呼ぶ。なお，連接類型の「連鎖型」については，普通は用いられない型であるとしているため，本研究では「連鎖型」を除いた7種を取り上げる。各連接類型の「接続語句のおもなもの」（pp.89-92）に挙げられている語句を下に示す。

順接型：だから，ですから，それで，したがって，そこで，そのため，そういうわけで，それなら，とすると，してみれば，では，すると，と，そうしたら，かくて， こうして，その結果，それには，そのためには

逆接型：しかし，けれども，だが，でも，が，といっても，だとしても，それなのに，しかるに，そのくせ，それにもかかわらず，ところが，それが

添加型：そして，そうして，ついで，つぎに，それから，そのうえ，それに，さらに，しかも，また，と同時に，そのとき，そこへ，次の瞬間

対比型：というより，むしろ，まして，いわんや，一方，他方，それに対し，逆に，かえって，そのかわり，それとも，あるいは，または

転換型：ところで，ときに，はなしかわって，やがて，そのうちに，さて，そもそも，いったい，それでは，では，ともあれ，それはそれとして

同列型：すなわち，つまり，要するに，換言すれば，言い換えれば，たとえば，現に，とりわけ，わけても，せめて，少なくとも

補足型：なぜなら，なんとなれば，というのは，ただし，もっとも，ただ，なお，ちなみに

[3] ただし，「されば」「じゃあ」「では」については，順接型と転換型の二種類の用法が存在する。今回のデータでは，どちらの連接類型か判断しにくい例が存在していたため，これらの接続詞は考察の対象から除外した。

[4] 本研究の接続詞は『中納言』の長単位検索による検索結果に依存している。そのため，接続詞の定義についての議論は行わない。

具体的には，接続詞の前一文字に文頭を表す記号（#）がある場合は，文頭として認定した。

さらに，分かち書き的に使用される読点を観察するために，文字種情報と読点の有無の関係について分析を行った。分析対象データから対象となる一文字（接続詞の初めの一文字と最後の一文字，接続詞の直前に書かれた単語の最後の一文字，接続詞の直後に書かれた単語の初めの一文字）を抽出し，それぞれに文字種情報を付与した。その後，「接続詞の直前に書かれた単語の最後の一文字 ＋ 接続詞の初めの一文字」の文字種の組み合わせ（以下，前文字種），「接続詞の後ろの一文字 ＋ 接続詞の直後に書かれた単語の最後の一文字」の文字種の組み合わせ（以下，後文字種）を作成して，それぞれの出現頻度と読点が打たれる確率を計算した（例(3)(4)では，文字の網掛け場所が前文字種，二重下線の所が後文字種である[5]）。たとえば，例(3) の前文字種は，「同機種」の「種」と接続詞「または」の「ま」であるため，「漢字_平仮名」となる。また，例(3) の後文字種は，接続詞「または」の「は」と「同等」の「同」であるため，「平仮名_漢字」となる。

なお，接続詞の直後に書かれた単語の初めの一文字が読点だった場合は，その次の一文字を文字種情報として使用している。例(4)で言えば，接続詞「一方」の「方」と「数日中」の「数」を取り上げ，後文字種は「漢字_漢字」となる。

(3) なんか、格安または無料で、番号そのままで同機種または同等の機種に変える方法ってありませんか？ 【出典】BCCWJ サンプル ID: OC02_03912 Yahoo!知恵袋

(4) ペルーの首都リマで二十日夜に発生した爆弾事件で、トレド大統領は二十二日、犯人逮捕につながる情報提供者に百万ドルの懸賞金を用意する一方、数日中に法改正を含むテロ対策を議会に提案する方針を明らかにした。
【出典】BCCWJ サンプル ID: PN2b_00004 毎日新聞 夕刊 2002/3/23

接続詞の前後に使用された品詞情報は，BCCWJ に付与されている情報を使用した。「中納言」による検索時に「ダウンロードオプション」から「インラインタグを使用」にチェックを入れ，品詞（大分類）タグを同時にダウンロードした。BCCWJ のコアデータは，自動解析後に人手修正が行われており，MeCab などによる自動解析よりも解析精度が高い（99%以上）高精度なデータである（小椋・冨士池,2015）。ただし，接続詞の直後に読点があるものに関しては，読点ではなくその後（つまり，接続詞から数えて二文字目）の形態素の品詞情報（例(4)で言えば，「数日」（名詞））を採用した。

４．分析

４.１　分析方法

今回の分析は，接続詞の直後に読点が打たれるか否かという，二つの値しか取り得ない二値データであるため，二項分布にしたがって分布することになる。そのため，一般化線形モデル（GLM）を用いたロジスティック回帰分析を行った[6]。ロジスティック回帰分析におけるモデル選択基準として，赤池情報基準（AIC）を採用した。AIC の値は「予測の悪さ」を

[5] 以下，全ての例文の下線や網掛けは筆者によるものである。

[6] 応答変数（読点の有無）は，読点ありを 1，読点なしを 0 と置き換えて分析をしている。つまり，とりうる値は{0, 1}である。そのため，glm 関数の family には binomial を選択した。なお，link 関数には，”logit”を使用している。

示し，この値が小さいほど，モデルの精度が高いと判定できるとされる（久保 2012)。このAIC に基づき，最も説明力の高い予測モデルを作成するために，今回はステップワイズ法（変数増加法）により変数選択を行った。

４.２　分析結果

４.２.１　ステップワイズ法による予測モデル（model1）

ステップワイズ法（変数増加法）で選択された予測モデル（model1）について紹介する。接続詞の直後の読点の有無を予測するモデルには，「接続詞の語彙素」，「レジスター」，「後文字種」，「前文字種」，「接続詞後の品詞」，「接続詞前の品詞」が選択された。その一方で，「一文の長さ」や「接続詞の文字数」，「接続詞の連接類型」，「接続詞の語種」，「文頭にあるかどうか」については，説明変数として扱われることがなかった。

model1 の AIC は 4406.2，残差逸脱度は 4190.2，残差自由度は 5439 であった。しかし，この model1 のピアソンの χ^2 統計量を計算したところ, 11950.09 であった。そのため, dispersion parameter の推定値（ピアソンの χ^2 統計量 / 残差自由度）は 2.197112 となり，過分散が起きていることが確認された。この過分散は説明変数内のパラメータが多すぎるために起きたと考えられる。そこで，パラメータが多い「接続詞の語彙素」と「レジスター」の説明変数を削除して，再度ステップワイズ法（変数増加法）を行った。

４.２.２　ステップワイズ法による変数を削除した予測モデル（model2）

上述した変数を削除して再度，予測モデルを構築した（model2)。表 2 は，model2 に対して行った anova（Type II）による各説明変数の逸脱度分析の結果である。

表 2 予測されたモデル（model2）の anova（TypeII）による結果

Step	Variable	LR Chisq	Df	Pr(>Chisq)	
+1	前文字種	317.39	13	5.49E-60	***
+2	後文字種	395.79	10	7.42E-79	***
+3	接続詞後の品詞	278.89	14	2.94E-51	***
+4	一文の長さ	122.08	1	2.21E-28	***
+5	接続詞の文字長	83.66	1	5.88E-20	***
+6	接続詞前の品詞	61.02	13	3.45E-08	***
+7	接続詞の連接類型	33.56	6	8.18E-06	***
+8	接続詞の語種	20.76	2	3.11E-05	***

---Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’

分析の結果, 接続詞の直後の読点の有無を予測するモデルには,「前文字種」,「後文字種」,「接続詞後の品詞」,「一文の長さ」,「接続詞の文字長」,「接続詞前の品詞」,「接続詞の連接類型」,「接続詞の語種」, が説明変数として扱われた。

model2 の AIC は 4900.7，残差逸脱度は 4776.7，残差自由度は 5485 であった。また，ピアソンの χ^2 統計量は 5645.093 であった。dispersion parameter の推定値（ピアソンの χ^2 統計量 / 残差自由度）は 1.029187 となり，過分散は起きていないことが確認された。

５．考察

表 3 は，model2 の説明変数の係数を示した表[7]である。説明変数を見てみると，まず，一文の長さが，最も影響があることがわかる。一文の長さが長くなるほど，読点が打たれやすくなっている。文が長いということは，読点の打つことができる場所も多くなる。つまり，文が長いほど読点の数が多くなる可能性があることを意味している。岩崎(2017 予定 a)では，他の品詞と比べた場合，接続詞の直後は読点が打たれにくいと指摘している。しかし，一文が長くなるほど，文の構造を示す上で必須の点以外の読点を打つことができると考えた場合，接続詞の直後に読点が打たれやすくなるという今回の分析結果は，岩崎(2017 予定 a)での指摘と整合性がある。

表 3 model2 の説明変数の係数

説明変数	Estimate	Std. Error	z value	Pr(>|z|)	RR95% CI .low	RR95% CI .up
一文の長さ	0.012	0.001	10.625	2.28E-26***	1.010	1.015
接続詞の文字数	-0.418	0.046	-9.055	1.37E-19***	0.601	0.721
連接類型:対比型	-1.041	0.256	-4.064	4.81E-05***	0.214	0.583
連接類型:同列型	-1.073	0.297	-3.609	0.000308***	0.191	0.612

---Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’

(5) が、ココはミクシィではないので、別に見たところでやらなくていいです（笑）。
【出典】BCCWJ サンプル ID:OY14_12862 Yahoo!ブログ

(6) すると瞳子さんはようやく頬を離して、目と鼻の下を手首でぐいっとぬぐった。
【出典】BCCWJ サンプル ID:PM11_00055 村山由佳(2001) 「ＩＮ　ＰＯＣＫＥＴ（月刊［文庫情報誌］)」

次に，接続詞自体の文字数については，文字数が長ければ長くなるほど読点が打たれにくいことを示している。その理由を考察すると，接続詞の成り立ちや見やすさが影響を与えていると思われる。たとえば，一文字しかない(5)の場合，見やすさ，接続詞であることをより目立たせるために読点を打った方がわかりやすい。また，元々の助詞としての成り立ちを考えれば，(5)の「が」は南（1974,1993）で言えば C 類にあたり，(6)の「と」は B 類に相当する。そのため，独立度が高くなればなるほど読点が打たれやすいという岩崎（2017 予定 b）の主張とも整合性がある。

そのほかにも，接続詞の連接類型から見た場合，対比型と同列型は読点が打たれにくいことが明らかになった。先述したように，これらの連接類型は文中のみに現れ，文頭では現れないものが多い連接類型であった。たとえば，例文(7)や(8)を見ると，接続詞「また」と「及び」は，名詞を列挙することで対比や同列を示しており，これらの例のような場合に読点が打たれる確率は極めて低いため，負の傾きを持つ説明変数としてあげられたと考えられる。

[7] 紙幅の都合上，有意差が確認されたものだけを提示している。

(7) なんか、格安または無料で、番号そのままで同機種または同等の機種に変える方法ってありませんか？

【出典】BCCWJ サンプル ID: OC02_03912 Yahoo!知恵袋

(8) 各年度ごとの調査対象企業数及び回答企業数は以下のとおり。

【出典】BCCWJ サンプル ID: OW6X_00054 経済産業省；厚生労働省；文部科学省(2004)『ものづくり白書 2004 年版』

６．まとめ

以上，一般化線形モデルによる分析の結果，接続詞後の読点の打たれやすさには，接続詞の意味機能（接続詞の連接類型）だけではなく，前後の文字種，接続詞前後の品詞，一文の長さ，接続詞の文字長，接続詞の語種といった要因が複雑に絡み合って存在していることが明らかになった。この結果は，石黒（2011）が指摘している読点が打たれる要因が，計量的な手法をもとにした分析結果から得られた要因とも一致する結果となった。

また，語彙素やレジスターを考慮した場合は，過分散が起こったため，読点の打たれやすさを予測するモデルには考慮しない方がいいことが判明した。これは，金・樺島・村上（1993）のように，判別には効果的である読点が予測には向かないという結果と一致している。一方では，レジスター間については差がないという小磯ほか（2008）の結果とも一致している。

当初は，文頭情報が接続詞の直後の読点の有無に影響を与えていると考えられたが，分析の結果は，文頭情報は説明変数として取り扱われなかった。これは，文頭情報が文頭に置かれるか否かといった二値しか持たないため，読点を打つかどうかという予測には不向きであったことが考えられる。

本研究で新たに明らかになったことを二点挙げる。一点目は，接続詞自体の文字長が読点の打ち方に影響を与えている点である。二点目は，表記（分かち書き）のために打たれる読点が，読点の有無の一般的な判別として扱うことができる点である。石黒（2011）では，この表記の読点は，個人差があるものとして取り上げているが，今回の分析では，読点の有無を判別する一つの説明変数として取り上げられた。

７．今後の課題

今後の課題としては，日本語教育や国語教育といった教育への示唆や，文章作成に対する提言を目標とした分析と考察を行いたい。今回の一般化線形モデルを用いた予測モデルは，どのような変数が接続詞直後の読点に影響を与えているのかを明らかにしたが，この予測モデルは，読点を打つか打たないかということを予測するだけであり，日本語（または国語）教育に役に立つモデルではない。そのため，今後は決定木分析を行うなど，どういう要因（判断基準）をもとに読点を打てばより適切になるのかといった教育への示唆を目的とした分析を行いたい。

もう一つの課題として，今回定義した予測モデルの精度の向上が挙げられる。他の説明変数を適宜取り入れたり削除したりするといったチューニングや内的／外的妥当性が存在するかどうかを検討する必要がある。今後は，BCCWJ のコアデータ以外のサンプルや他の文章データを用いた場合にも当てはまりが良いかどうか判別分析などを用いて予測モデルの精度の向上を試みる。

さらに，読点の打ち方が書き手による個性であるのならば，これら個人の癖を考慮した分析を行う必要がある。今回は，一文中における読点の数や段落（文章）構造などといった複雑な変数は取り扱わなかった。また，今回過分散を起こしたために説明変数から外した語彙素やレジスターといった変数については，ランダム効果として組み込み，一般化線形混合モデルを用いることで予測モデルのさらなる精緻化を目指したい。

文　献

石黒圭（2011）.「14 句読点のルール」中村明・佐久間まゆみ・髙崎みどり・十重田裕一・半沢幹一・宗像和重(編)『日本語文章・文体・表現事典』, pp.301-304, 朝倉書店.

市川孝（1978）.『国語教育のための文章論概説』, 教育出版.

岩崎拓也（2016）.「中国人・韓国人日本語学習者の作文に見られる句読点の多寡」『一橋日本語教育研究』, 第 4 号, pp187-196, ココ出版.

岩崎拓也（2017 予定 a）.「日本語学習者の作文コーパスから見た読点と助詞の関係性」『一橋大学国際教育センター紀要』, 第 8 号, 一橋大学国際教育センター.

岩崎拓也（2017 予定 b）.「第 5 章 正確で自然な句読点の打ち方」石黒圭(編)『わかりやすく書ける作文シラバス』, くろしお出版.

小椋秀樹・冨士池優美（2015）.「第 5 章 形態論情報」『『現代日本語書き言葉均衡コーパス』 利用の手引　第 1.1 版』, pp58-90, 国立国語研究所 コーパス開発センター.

金明哲・樺島忠夫・村上征勝（1993）.「読点と書き手の個性」『計量国語学』18(8), pp.382-391, 計量国語学会.

久保拓弥（2012）.『データ解析のための統計モデリング入門: 一般化線形モデル・階層ベイズモデル・MCMC』, 岩波書店.

小磯花絵・小木曽智信・小椋秀樹・冨士池優美・宮内佐夜香（2008）.「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』, pp.192-195, 社会言語科学会.

南不二男（1974）.『現代日本語の構造』, 大修館書店.

南不二男（1993）.『現代日本語文法の輪郭』, 大修館書店.

文部省（1963）.「付録 句切り符号の使ひ方〔句読法〕(案)」『(国語シリーズ 56) 国語表記の問題』, pp.60-76,教育図書.
(http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/series/56/pdf/kokugo_series_056_05.pdf :2017 年 6 月 11 日アクセス)

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/