連濁に前部要素の音韻的特徴が与える影響：連濁データベースを利用した研究

太田　真理　（東京大学大学院総合文化研究科）
太田　聡　（山口大学人文学部）

Effects of First-Element Phonological Features on Rendaku: A Study Using the Rendaku Database

Shinri Ohta (Graduate School of Arts and Sciences, The University of Tokyo)
Satoshi Ohta (Faculty of Humanities, Yamaguchi University)

要旨

本研究では、連濁データベースに含まれる複合名詞のうち、先行研究で提案された音韻・統語・意味的要因からは連濁を予測できない 16,211 語に対して、前部要素と後部要素が持つ子音、母音、アクセント、モーラ数などの音韻的特徴から、連濁の生起が予測可能かどうかを検証した。これらの例外とされる複合語においても、後部要素の音韻的特徴に加えて前部要素の音韻的特徴を考慮することで、連濁の生起が予測可能である、という仮説を立てた。サポートベクターマシン (SVM) を用いた解析の結果、後部要素のみに基づくモデルの正答率は、チャンスレベルより有意に高く、さらに前部要素の音韻的特徴を加えると 90%以上の複合語で正しく連濁の生起を予測可能であった。以上の結果から、連濁現象の例外とされてきた複合語でも、後部要素の音韻情報に加えて、前部要素の音韻情報を考慮に入れることで、連濁の生起が予測可能であることが示された。

1. 日本語の連濁現象

1.2. 本居・ライマンの法則

日本語では、複合語の後部要素が、有声性に関して対立を持つ無声阻害音（清音：/k/, /s/, /t/, /h/）で始まる場合に、対応する有声阻害音（濁音：/g/, /z/, /d/, /b/）に変化する、連濁現象が知られている。例えば、「ごみ＋はこ」は「ごみばこ」となり（下線部は連濁を示す）、後部要素「はこ」が清音/h/で始まるために、対応する濁音/b/に変化する。一方で、「ごみ＋かご」が「ごみがご」にならないことから明らかなように、連濁は常に生じるわけではない。連濁を阻害する制約として、「複合語の後部要素が濁音を含む場合は連濁が生じない」という「本居・ライマンの法則」が知られている (Lyman, 1894)。この法則から、「ごみ＋はこ」は「はこ」が濁音を含まないために連濁が生じて「ごみばこ」となるが、「ごみ＋かご」では「かご」が濁音/g/を含むために連濁が妨げられて「ごみかご」となることが正しく予測される。

1.2. 連濁に影響する形態・統語・意味的要因

本居・ライマンの法則に加えて、単語の形態論的要因や統語的要因、意味的要因が連濁の生起に影響を与えることが示唆されてきた。例えば、形態論的要因として、「ごみ＋ケース」が「ごみゲース」とならないように、連濁は主に和語で観察され、漢語や外来語では観察されないという語種の効果が知られている（「会社」、「キセル」などの一部の和語化した単語では例外的に連濁が生じる）(Ito and Mester, 2003)。また、統語的要因として、

「ぬりばしいれ（塗り箸専用の入れ物）」と「ぬりはしいれ（漆塗りの箸入れ）」の対比から示されるように、「複合語の木構造中で右枝に来る要素のみで連濁が生じる」という、「右枝条件」が知られている (Otsu, 1980)。

意味的要因として、「やま＋かわ」は「やまかわ」となるが、「たに＋かわ」は「たにがわ」となることから分かるように、前部要素と後部要素が意味的に並列される複合語（並列複合語）では、連濁が妨げられることが知られている (Ito and Mester, 2003)。一方、「ひと＋ひと」が「ひとびと」となるように、同一の単語の繰り返しからなる複合語（畳語）では、連濁が生じる。さらに、「やま＋さき」が「やまざき」にも「やまさき」にもなりうるように、人名や地名などの固有名詞では、連濁の生起に関して曖昧性が生じやすいことも知られている。

1.3. 前部要素が連濁に与える影響

ここまで挙げた要因は、後部要素に含まれる要因（例えば本居・ライマンの法則や語種の効果）、あるいは、前部要素と後部要素の関係によって決まる要因（右枝条件や並列複合語）であった。これに対して、「なかじま」と「ながしま」の対立から示唆されるように、「複合語の前部要素が濁音を含む場合も連濁が生じない」という「強いライマンの法則」が提案されている (Vance, 2005)。この強いライマンの法則は上代日本語では機能していたと考えられるが、現代日本語でも機能しているかどうかについてははっきりしていない (Zamma, 2005; Kawahara and Sano, 2014)。連濁の生起に関してさまざまな要因が影響することを概観したが、いずれの要因にも例外があり、完全に連濁の生起を説明できるわけではない。このように例外が多い連濁現象から一般則を導くためには、多次元の情報に基づいて、統計的に結果を予測する機械学習を用いることが有効であると考えられる。

1.4. 研究の目的

本研究では、連濁データベース (Miyashita and Irwin, 2014) に含まれる日本語複合名詞の中で、音韻・形態・統語・意味的な要因では連濁の生起が説明できない語を対象に、前部要素と後部要素の音韻的特徴（子音・母音・モーラ数・アクセント）の組み合わせによって、連濁の生起が予測できるかどうか検証することを目的とした。我々は、これらの例外として扱われてきた複合語においても、後部要素の音韻的特徴に加えて、前部要素の音韻的特徴も考慮に入れることで、連濁の生起が予測可能である、という仮説を立てた。この仮説を検証するために、統計的機械学習の分野の標準的な手法であるサポートベクターマシン (SVM) を用いて、音韻的特徴から連濁を予測した際の正答率を調べた。

2. 研究方法

2.1. 連濁データベース

本研究で使用した連濁データベースは、国立国語研究所共同研究プロジェクト「日本語レキシコン―連濁事典の編纂」の一環として構築が進められており、本研究ではその最新バージョンである v2.3 を使用した (Miyashita and Irwin, 2014)。連濁データベースには、広辞苑または新和英大辞典に含まれる 32,241 個の複合語が収録されており、これら複合語は以下のいずれかに該当する。

(1) 後部要素が和語で、本居・ライマンの法則によって連濁が阻害されない複合語
(2) 後部要素が漢語または外来語で、例外的に連濁が生じる複合語
(3) 「はしご」のように、ライマンの法則に反して連濁が生じる複合語

このデータベースには、複合語ごとの連濁の有無に加えて、前部要素と後部要素の語種、モーラ数、品詞、さらに後部要素の使用頻度、アクセントなどの情報も記載されている。

本研究では、連濁データベースに含まれる複合語のうち、後部要素が和語の名詞であり、連濁の生起に曖昧性が存在しない複合語を解析の対象とした。また、先行研究から示唆された連濁に影響する音韻・統語・意味の要素を排除するために、以下の基準に該当する複合語は解析対象から除外した。

a. 人名・地名でのみ使用される単語
b. 並列複合語で連濁が生じない単語
c. 畳語で連濁が生じる単語
d. 省略語
e. /b/に由来する/m/を含み、連濁が生じない単語

16,211 個の複合語を解析の対象とし、このうち 13,115 個で連濁が生起し、連濁の生起率は 80.9%であった。

2.2. サポートベクターマシンを利用した連濁生起の予測

SVM とは、機械学習で使われる手法の一つであり、特にパターン認識に関して優秀な学習モデルであることが知られている。図 1 では、○と●を分類する場合を例に SVM の説明を行う。○と●を分ける分離超平面の引き方は無数に存在するが、SVM ではマージン（分離超平面とデータとの距離）が最大となるように、分離超平面を決定するする。この時の分離超平面と最も近いデータのことをサポートベクトルと呼ぶ。

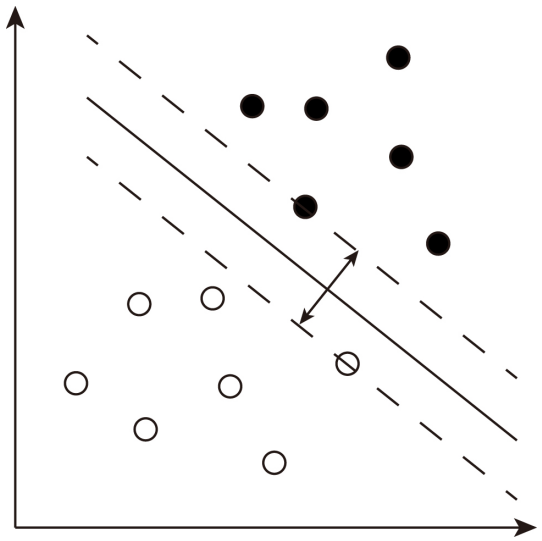

図 1. サポートベクターマシンの概略
実線は分離超平面を示し、矢印はマージンを示す。SVM はマージンが最大となる分離超平面を学習する（破線上のデータはサポートベクトル）。

音韻的特徴から連濁の生起が予測可能かどうかを SVM により検証するために、以下のように単語の音韻情報を 2 値的にコード化して、モデルに取り入れた。まず、前部要素の語末の音節と、後部要素の語頭の音節を取り出した。これらの音節をそれぞれ子音と母音に分け、個々の音素を音韻的特徴として抽出した。例えば「ごみ＋はこ」という複合語において、前部要素の最後のモーラは/m/, /i/という音素を含み、後部要素は/h/, /a/という音素を含むため、これらの音素を 1 とし、それ以外の音素は 0 とする。同様に、前部要素全体に含まれる音素についてもモデルに組み込んだ。例えば前部要素が「やま」の場合、/y/, /m/, /a/という音素を持つ。ここで、/a/は単語中に 2 個含まれるため、/y/, /m/, /a/はそれぞれ 1, 1, 2 とする。

さらに、前部要素のモーラ数、後部要素のモーラ数、複合語全体のモーラ数、後部要素のアクセント（現代語でのアクセントと古語でのアクセント）についての情報も学習に利用した。以上の結果、前部要素に含まれる子音・母音に関する特徴 26 種類、前部要素の最後の音節に関する特徴 22 種類、後部要素の最初の音節に関する特徴 10 種類、モーラ数に関する特徴 3 種類、アクセントに関する特徴 2 種類、使用頻度に関する特徴 1 種類の全 64 種類の情報を利用した。SVM の学習ではデータのスケールをそろえる必要があるため、個々の音韻的特徴 x に対して、

$$\{x - \min(x)\}/\{\max(x) - \min(x)\}$$　（$\max(x)$ は x の最大値、$\min(x)$ は x の最小値）

を適用して、x が 0 から 1 の間の値を持つように調整した (Hsu et al., 2003)。

SVM を使ったデータ解析には、統計言語 R の e1071 パッケージに含まれる svm 関数を使用した (Meyer et al., 2014)。また、学習データに対するオーバーフィッティングを避けるために、50 分割交差確認を行った。50 分割交差確認とは、データを 50 個に分割し、そのうちの 49 個を使用して学習を行い、使用していないデータで SVM の検証を行う、というプロセスを 50 個のデータそれぞれに対して行う手法である。SVM による予測がチャンスレベルよりも高いかを t 検定により調べた。解析対象にした複合語では、81%で連濁が生じていたため、81%に比べて有意に正答率が高い場合、SVM による学習は成功したと考えられる。まず、後部要素の音韻的特徴のみに基づいて連濁を予測するモデルが、連濁の生起をどの程度予測できるかを調べ、さらにこのモデルに前部要素の音韻的情報を加えた場合にどの程度予測精度が向上するかを検証した。最後に、前部要素の音韻的特徴のみに基づいて連濁を予測するモデルも構築し、モデル間の正答率を分散分析により比較した。

3. 結果

3.1. 前部要素と後部要素の音韻的特徴を組み合わせることで説明可能な連濁

後部要素の音韻的特徴のみに基づくモデルの正答率は 83 ± 1.8 %であり、チャンスレベル（81%）よりも有意に高い正答率であった ($p < 0.0001$, $t(49) = 7.5$)。これに対して、後部要素の音韻的特徴に前部要素の音韻的特徴を加えたモデルは、90 ± 1.4 %のデータに対して正しく連濁の生起を予測した（図 2）。このモデルの正答率も、チャンスレベルよりも有意に高かった ($p < 0.0001$, $t(49) = 46$)。これに対して、前部要素の音韻的特徴のみに基づくモデルの正答率は 81 ± 2.2 %であった ($p = 0.58$, $t(49) = 0.56$)。

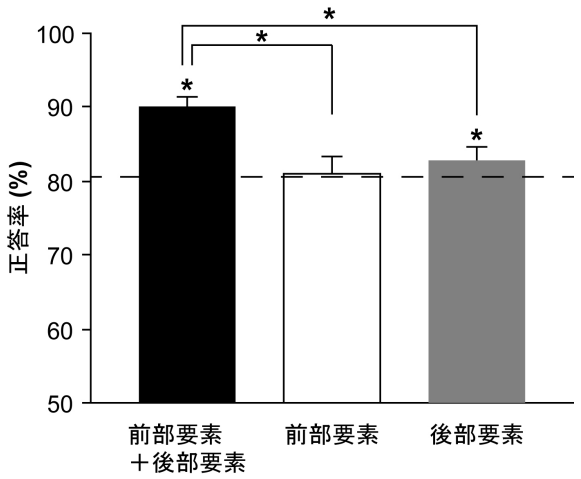

図 2. 前部要素と後部要素の音韻的特徴に基づく連濁の予測

各モデルの正答率を示す（平均±標準偏差）。破線はチャンスレベルを示す。前部要素＋後部要素：前部要素と後部要素の音韻的特徴を組み合わせたモデル。前部要素：前部要素の音韻的特徴のみに基づくモデル。後部要素：後部要素の音韻的特徴のみに基づくモデル。
*：corrected $p < 0.05$。

3.2. モデル同士の正答率の比較

これらの 3 種類のモデルの正答率に対して、繰り返しのない分散分析を行った結果、有意なモデルの主効果が観察された ($p < 0.0001$, $F(2,147) = 329$)。対応のない t 検定により、さらにモデル同士の正答率を比較したところ、前部要素と後部要素の音韻的特徴を組み合わせたモデルは、他のモデルよりも有意に正答率が高かった ($p < 0.0001$, $t(98) > 22$)。また、後部要素の音韻的特徴のみに基づくモデルは、前部要素の音韻的特徴のみに基づくモデルよりも有意に正答率が高かった ($p < 0.0001$, $t(98) = 4.3$)。以上の結果から、後部要素の音韻的特徴から、ある程度連濁の生起が予測可能であり、さらに前部要素の音韻的特徴を加えることで、非常に高い精度で連濁の生起が予測可能であることが明らかとなった。

4. 考察

本研究では、前部要素と後部要素の音韻的特徴から連濁の生起が正しく予測されるかどうかを、機械学習の手法である SVM により検証した。連濁データベースに含まれる日本語複合名詞のうち、先行研究で提案された音韻・統語・意味的要因からは、連濁が予測できない 16,211 個の複合語を対象とした。後部要素の音韻的特徴のみに基づいて連濁を予測するモデル、後部要素の音韻的特徴に加えて前部要素の音韻的特徴も考慮して連濁を予測するモデル、前部要素の音韻的特徴のみに基づいて連濁を予測するモデル、という 3 種類のモデルの比較検討を行った。後部要素の音韻的特徴のみに基づくモデルでは、83%の複合語に対して、正しく連濁の生起を予測可能であった。しかしながら、今回使用したデータのうち、81%の複合語では連濁が生じていたため、このモデルでは十分に予測精度が改善したとは言い難い。これに対して、前部要素と後部要素の音韻的特徴に基づくモデルでは、90%以上の高い精度で連濁の生起を予測可能であった。これらの結果は、連濁現象の例外とされてきた複合語では、後部要素の音韻的特徴に加えて、前部要素の音韻的特徴も考慮することが、正しく連濁の生起を予測するために必須であることを示唆する。一方

で、前部要素の音韻的情報のみに基づくモデルでは、チャンスレベルと比べて有意に予測精度が向上しなかったことから、現代日本語では「強いライマンの法則」は機能していないことが示唆される。また本研究は、このように多くの要因が関係する言語現象にとって、統計的機械学習の手法が極めて有効であることを示すものである。

本研究で用いたSVMは、与えた学習データに対して1つの分離超平面を学習する手法であり、学習データ中のどの情報が、予測力の向上に重要であったのかは明らかでない。今回考慮した音韻的特徴には、子音・母音・アクセント・モーラ数という性質の異なる特徴が混在しており、それぞれ連濁の予測に対する寄与率が異なることが予想される。今後の研究では、判別分析やロジスティック回帰分析のように、個々の説明変数に対して予測への寄与率が計算される統計手法を用いることで、重要度の高い音韻的特徴の絞り込みを行う予定である。また、今回は考慮に入れなかった統語・意味的要因や、前部要素のアクセント等をモデルに組み込むことで、さらなる予測精度の向上を目指す予定である。

謝　辞

本研究で使用した「連濁データベース」をご提供くださった山形大学のアーウィン先生並びにモンタナ大学の宮下先生に感謝いたします。

文　献

Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin (2003) “A Practical Guide to Support Vector Classification,” http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Ito, Junko and Armin Mester (2003) *Japanese Morphophonemics: Markedness and Word Structure*, Linguistic Inquiry Monograph 41, Cambridge, MA: The MIT Press.

Kawahara, Shigeto and Shin-ichiro Sano (2014) “Testing Rosen’s Rule and Strong Lyman’s Law,” *NINJAL Research Papers*, 7, pp. 111–120.

Lyman, Benjamin S. (1894) “Change from surd to sonant in Japanese compounds,” *Oriental Studies of the Oriental Club of Philadelphia*, pp. 1–17.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, and Chih-Chen Lin (2014) “Package ‘e1071,’” http://cran.r-project.org/web/packages/e1071/e1071.pdf.

Miyashita, Mizuki and Mark Irmin (2014) The Rendaku Database v2.3 (http://www-h.yamagata-u.ac.jp/~irwin/site/Rendaku_Database.html よりダウンロード可能)

Otsu, Yukio (1980) “Some Aspects of Rendaku in Japanese and Related Problems,” *MIT Working Papers in Linguistics*, 2, pp.207–227.

Vance, Timothy J. (2005) “Sequential Voicing and Lyman’s Law in Old Japanese.” In Salikoko S. Mufwene, Elaine J. Francis & Rebecca S. Wheeler (eds.), *Polymorphous linguistics: Jim McCawley’s legacy*, pp. 27–43. Cambridge: The MIT Press.

Zamma, Hideki (2005) “Correlation between Accentuation and Rendaku in Japanese Surnames: With Particular Attention to Morphemes.” In Jeroen van der Weijer, Kensuke Nanjo & Tetsuo Nishihara (eds.), *Voicing in Japanese*, pp. 157–176. Berlin: Mouton De Gruyter.