# 連濁の形式的アナロジーモデル

淺尾仁彦 (京都大学大学院)
asaokitan@ling.bun.kyoto-u.ac.jp

## 1 はじめに

日本語の連濁現象は，ライマンの法則[1]をはじめとして，その生起にかかわる条件について音韻・形態・意味面にわたってさまざまな一般化が提案されている (佐藤 1989) が，例外も多く，完全な予測は困難であることが知られている。

本発表では，部分的規則性と例外の多く見られる現象の予測に適した memory-based learning (Daelemans et al. 2005, 2007) を日本語の連濁現象に適用し，アナロジーが連濁を高い確度で予測すること，特に後部要素を共通してもつ語からのアナロジーが 9 割の語で連濁を正しく予測することを示す。

## 2 先行研究

### 2.1 連濁とアナロジー

連濁において既知の類似した語からのアナロジーが働いていることを支持する実験研究に Vance (1980), Ohno (2000) がある。Vance (1980) は，未知の語に対する話者のふるまいは，既存の語 (共通の後部要素をもつ語) の連濁パタンに従っていると見なせば十分であり，ライマンの法則の心理的実在性は弱いと論じている。また Ohno (2000) は，後部要素が共通する語のなかでも特に，意味的ないし音韻的に類似した特定の語からのアナロジーが働くことを示している。

連濁におけるアナロジーはこれまで，規則 (制約) に基づくアプローチ (e.g. Ito & Mester (2003)) ほど形式化されていないが，既知語の集合と類似性の定義を明示的に与えることにより，アナロジーにも予測力を持たせることが可能である。

### 2.2 オランダ語の linking morpheme との並行性

連濁のアナロジーモデルを形式化するうえで，オランダ語の linking morpheme の研究が参考となる。

linking morpheme は，複合語形成において 2 つの構成素の間に現れることがある *-s-* [s]，*-en-* [ə] という要素を指す。

(1) a. schaap-herder ‘shepherd’
b. schaap-s-kooi ‘sheepfold’
c. schap-en-vlees ‘mutton’
(Baayen 2003: 243)

ある複合語にどの linking morpheme が用いられるかは，生産的であるにもかかわらず予測が難しいことが知られている。Krott et al. (2002: 182) によれば，(i) full vowel のあとでは linking morpheme は現れない，(ii) 名詞化接辞 *-heid* のあとでは linking morpheme は *-s-*，(iii) 前部要素が後部要素の目的語ならば linking morpheme は現れない[2]など，さまざまな一般化が提案されているが，その多くに例外がある。Krott らによれば，提案されている音韻的・形態的規則は，Krott らのデータのうち 32 % の複合語についてしか正しい予測を行わない。

これに対し，Krott et al. (2001, 2002) は形式的アナロジーモデルが約 92 % の複合語を正しく予測することを示した。

van de Weijer (2003) は，オランダ語の linking morpheme と日本語の連濁の並行性を指摘し，linking morpheme の分析を連濁に適用することの有意義性を示唆している。そこで，本発表ではこのアナロジーモデルを日本語の連濁に適用することを試みる。

[1] 複合語後部要素が濁音 (有声阻害音) を含む場合は連濁を生じない。

[2] この条件は日本語の動詞由来複合語 (deverbal compound) の連濁について知られている一般化と同一であり興味深い。

## 3　連濁のアナロジーモデル：方法

アナロジーモデルでは，言語知識は用例の大規模な記憶からなり，未知語のふるまいは，抽象化された規則によってではなく，類似した既知語を直接参照することによって決まると考える。

本発表では，具体的なアルゴリズムとして Tilburg Memory-Based Learner (TiMBL) (Daelemans et al. 2005, 2007) を用いた。TiMBL では，各語彙を素性の束によって表現し，ある語に連濁が生じるかどうかは，素性の一致度が最も高い (= 最近傍の) 既知語で連濁が生じているかどうかによって決まる[3]。

### 3.1　データ

本発表ではデータとして『日本語の語彙特性』(天野・近藤 1999) を利用し，後部要素の初頭に /k, s, t, h/ のいずれかをもつ複合語を機械的に抽出した。今回は，複合語およびその構成要素がともに名詞・動詞・形容詞・形容動詞のいずれかに同定できるものに対象を限定した。また，カタカナ語は除外し，ひらがなのみで表記された語も，複合語かどうかの判断が難しいためデータから除いた。明らかな分析誤りの例を排除し，表 1 の 10,047 語を得た。

| | 後部要素が和語 | 後部要素に漢語を含む | 計 |
|---|---|---|---|
| 連濁なし | 2,565 | 4,518 | 7,083 |
| 連濁あり | 2,584 | 380 | 2,964 |
| 計 | 5,149 | 4,898 | 10,047 |

表 1　対象とした複合語の総数

### 3.2　素性

次に，前節で抽出した 10,047 語の複合語について，以下の 25 の素性により表示した。

まず単純に，同一の構成素をもつ既知語からのアナロジーを行っている可能性を考え，前部要素，後部要素の語自体を素性として立てた。なお，「-塩」(連濁あり) と「-潮」(連濁なし) のように，連濁が同音語の区別に貢献している可能性が指摘されていることから (Vance 1987: 147)，漢字表記を含めてデータとして与えた。次に，ライマンの法則など語を構成する音素が連濁の生起に影響している可能性を考え，各音素を素性として立てた。複合動詞は連濁しないこと[4]，前部要素が形容詞由来であるとき連濁が起きにくいこと[5]などが指摘されていることから，品詞に関する情報を素性として立てた。また，和語と漢語との区別を素性として立てた。さらに，モーラ数が連濁の生起と相関する場合が指摘されていることから (Ito & Mester (2003: 275), Ohno (2000: 161))，モーラ数を素性として立てた。まとめると以下の通りである[6]。

(2)　a. 前部要素 ($f_1$)
　　b. 後部要素 ($f_2$)
　　c. 前部要素と後部要素の境界から前後それぞれ 4 モーラの子音および母音 ($f_3$〜$f_{18}$)
　　d. 前部要素の品詞 ($f_{19}$)
　　e. 後部要素の品詞 ($f_{20}$)
　　f. 複合語の品詞 ($f_{21}$)
　　g. 前部要素の語種 ($f_{22}$)
　　h. 後部要素の語種 ($f_{23}$)
　　i. 前部要素のモーラ数 ($f_{24}$)
　　j. 後部要素のモーラ数 ($f_{25}$)

[3] 最近傍に語が複数ある場合は多数決となり，同数で決着がつかない場合は二番目に近い語まで含めて同じプロセスが繰り返される。

[4] 複合動詞 943 語中，連濁した語は 7 語のみであった。

[5] 佐藤 (1989) は次のような例を挙げている:

(i)　a. 暑苦しい，狭苦しい (くるしい)
　　息苦しい，心苦しい (ぐるしい)
　　b. 甘口，薄口 (くち)
　　戸口，告げ口 (ぐち)

「-苦しい」，「-口」に関しては確かに形容詞のあとで濁らないという現象が見られるようである。しかしながら今回のデータ全体 (後部要素が和語のもの) を見ると，前部要素が形容詞の語は連濁なしが 119 語，連濁ありが 182 語で，予想に反して，有意に連濁を起こしやすい ($\chi^2 = 13.1, p < .001$) という結果が出た (ただし，データから複合動詞を除くと有意差は見られなかった)。

[6] 連濁に関してよく知られた一般化のうち，今回立てた素性では捉えられないと思われるものも多い。例えば，右枝分かれ制約 (Otsu 1980)，並列語 (dvandva) が連濁しないこと，動詞由来複合語のうち内項の複合した語は連濁しにくいこと (定量的記述が 高野 (2006) にある)，語彙的アクセントとの相関などがある。

| ID | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $f_7$ | $f_8$ | $f_9$ | $f_{10}$ | $f_{11}$ | $f_{12}$ | $f_{13}$ | $f_{14}$ | $f_{15}$ | $f_{16}$ | $f_{17}$ | $f_{18}$ | $f_{19}$ | $f_{20}$ | $f_{21}$ | $f_{22}$ | $f_{23}$ | $f_{24}$ | $f_{25}$ | 連濁? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 095260 | 紙 | 漉き | = | = | = | = | k | a | m | i | s | u | k | i | = | = | = | = | 名 | 動 | 名 | 和 | 和 | 2 | 2 | No |
| 095300 | 神 | 頼み | = | = | = | = | k | a | m | i | t | a | n | o | m | i | = | = | 名 | 動 | 名 | 和 | 和 | 2 | 3 | Yes |
| 095330 | 噛み | 付く | = | = | = | = | k | a | m | i | t | u | k | u | = | = | = | = | 動 | 動 | 動 | 和 | 和 | 2 | 2 | No |
| 095340 | 紙 | 包み | = | = | = | = | k | a | m | i | t | u | t | u | m | i | = | = | 名 | 動 | 名 | 和 | 和 | 2 | 3 | Yes |
| 095360 | 噛み | 潰す | = | = | = | = | k | a | m | i | t | u | b | u | s | u | = | = | 動 | 動 | 動 | 和 | 和 | 2 | 3 | No |
| 095370 | 紙 | 礫 | = | = | = | = | k | a | m | i | t | u | b | u | t | e | = | = | 名 | 名 | 名 | 和 | 和 | 2 | 3 | No |
| 095390 | 上 | 手 | = | = | = | = | k | a | m | i | t | e | = | = | = | = | = | = | 名 | 名 | 名 | 和 | 和 | 2 | 1 | No |
| 095400 | 紙 | 鉄砲 | = | = | = | = | k | a | m | i | t | e | - | Q | p | o | - | u | 名 | 名 | 名 | 和 | 漢 | 2 | 4 | Yes |
| 095410 | 髪 | 床 | = | = | = | = | k | a | m | i | t | o | k | o | = | = | = | = | 名 | 名 | 名 | 和 | 和 | 2 | 2 | Yes |
| 095480 | 紙 | 挟み | = | = | = | = | k | a | m | i | h | a | s | a | m | i | = | = | 名 | 動 | 名 | 和 | 和 | 2 | 3 | Yes |

表 2　データの例 (一部簡略化)

素性によって表示したデータの例を表 2 に示した。

### 3.3　類似度の計算方法

類似度の計算方法に関しては様々な手法が考えられるため，いくつかの手法の比較を行った。まず，類似度の尺度として，Overlap と modified value difference (MVDM) の 2 つを比較した。Overlap 法は単純に，それぞれの素性について値が一致しているかいないかのみを問題にするのに対し，MVDM は，あらかじめ素性の各値について連濁の条件付確率を求めることで，b と d の類似度は b と s の類似度よりも高いといったことを反映させる方法である。

また，素性ごとに連濁における重要度が異なることを考慮し，Gain Ratio (Quinlan 1993) [7]による素性の重み付けを行った場合と，重み付けをしない場合について比較をした。

leave-one-out 交差検定法によってそれぞれの方法の成績を調べたところ，表 3 の結果が得られた[8]。

表 3 のように，MVDM の利用は必ずしも成績の改善をもたらさなかった。一方，素性の重み付けを行うことで，Overlap 法におけるパフォーマンスは 87.2 % から 90.3 % に有意に改善した ($\chi^2 = 47.8, p < .001$)。

| 方法 | 正解数 | 正解率 (%) |
|---|---|---|
| Overlap + 重み付けなし | 8,759 | 87.2 |
| Overlap + Gain Ratio | 9,070 | 90.3 |
| MVDM + 重み付けなし | 8,891 | 88.5 |
| MVDM + Gain Ratio | 8,967 | 89.3 |

表 3　方法別の成績 (類似度尺度と素性重み付け)

また，以上は最近傍の語だけをアナロジーのベースとして用いているが，2 番目以降に近い語も参照する方法もある。Overlap + Gain Ratio の条件で，最近傍の語だけを用いた場合 ($k = 1$) のほか，5 番目に近い語まで含めて多数決を行った場合 ($k = 5$)，またさらに inverse distance 法によって，より近い語のふるまいが重視されるよう重み付けした場合を比較した。結果は表 4 のようになり，有意な改善は見られなかった。

| 方法 | 正解数 | 正解率 (%) |
|---|---|---|
| $k = 1$ | 9,070 | 90.3 |
| $k = 5$ | 9,040 | 90.0 |
| $k = 5$ (inverse distance) | 9,091 | 90.5 |

表 4　方法別の成績 (参照する近傍の数)

以上を踏まえ，次節では最も成績の良かった Overlap + Gain Ratio の組み合わせを用い，$k = 1$ として分析を行う。

[7] Gain Ratio は，ある素性の値を得たときに連濁について得られる平均の情報量 (エントロピーの減少分) である (ただし，値の種類の多い素性が有利になるのを防ぐため，素性自体のエントロピーを用いて補正してある)。Gain Ratio が大きい素性ほど，連濁の生起に大きく関与しているといえる。

[8] 素性 $f_1$, $f_2$ はデータが著しく疎 (sparse) であり MVDM に適さないため常に Overlap 法を用いた。なお，$f_{24}$, $f_{25}$ は数値データのため，数値の近さが直接反映される別の尺度を試みたが，成績の改善は見られなかった。

## 4　連濁のアナロジーモデル：結果と考察

### 4.1　素性の重み

まずここでは，前節で利用した各素性の Gain Ratio を見る。Gain Ratio の高い (従って，連濁への影響力が強い) 素性は，順に $f_{23}$, $f_{22}$, $f_2$, $f_{20}$, $f_{13}$ であった (図 1)。上位 2 つの $f_{23}$, $f_{22}$ は語種の素性であり，漢語・和語の区別が連濁の生じる可能性に大きく関与していることがわかる。5 位の $f_{13}$ は後部要素 2 モーラめの子音であり，ここに濁音が現れた場合は，連濁が起きない確実な証拠となる (今回のデータのなかに例外は存在しなかった)。つまり，モデルがライマンの法則のパタンに反応したものとみることができる。

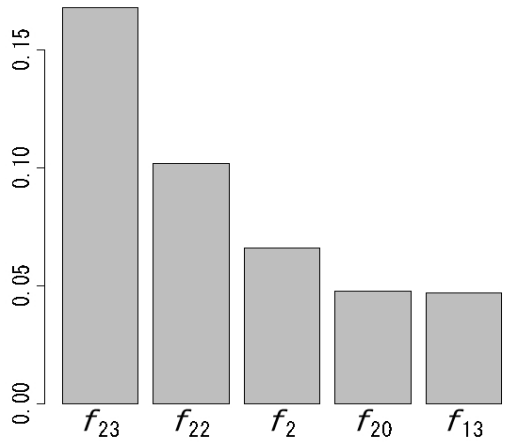


図 1　素性の Gain Ratio (上位 5 つ)

### 4.2　素性の絞り込み：パタンの検証

既に見たように，25 個の素性を用いることで 90.3 % の語の連濁を予測することができた。ここでは用いる素性を絞り込むことで，先行研究で指摘されたアナロジーのパタンを確かめる[9]。

表 5 に，いくつかの素性だけを用いた場合の結果を示した。

| 利用した素性 | 正解数 | 正解率 (%) |
|---|---|---|
| ベースライン[10] | 7,083 | 70.5 |
| 語種 ($f_{22}, f_{23}$) | 7,296 | 72.6 |
| 品詞 ($f_{19}$〜$f_{21}$) | 7,743 | 77.1 |
| 後部要素の音素 ($f_{11}$〜$f_{18}$) | 8,776 | 87.3 |
| 後部要素 ($f_2$) | 9,039 | 90.0 |
| 全素性 | 9,070 | 90.3 |

表 5　素性別の成績

表 5 から，後部要素 ($f_2$) を唯一の素性として用いても 90.0 % の正解率を示すことがわかる[11]。このことは，連濁は後部要素を共通してもつ語に合わせてふるまっているとする Vance (1980) の議論と一致する。

また，Krott et al. (2001: 73) は，素性として前部要素のみを用いることでオランダ語の linking morpheme の 92.5 % を正しく予測している。van de Weijer (2003: 272) が述べる通り，オランダ語の linking morpheme では前部要素の共通性が，日本語の連濁では後部要素の共通性がアナロジーの最も強力な手がかりとして機能することがわかる。

後部要素が和語である 5,149 語に対象を限定した場合の結果は表 6 のようになった。

表 6 のように，後部要素 ($f_2$) のみを素性として用いた場合の成績は 81.9 % であったが，品詞や音素の素性を加えた場合はこれより有意に高い成績を示した (品詞の場合 $\chi^2 = 43.1, p < .001$，音素の場合 $\chi^2 = 46.4, p < .001$)。これは，単に後部要素を共通してもつ語だけでなく，音素や品詞の一致をアナロジーの際に利用することが有用であることを意味している。

[9] TiMBL では，素性の重み付けを行っている場合でも，不要な素性の追加は成績をかえって低下させることがある (Daelemans et al. 2007: 22)。この現象は表 6, 7 に現れている。

[10] 全て連濁しないとした場合の成績。

[11] 参考に，アナロジーを用いず (i) ライマンの法則 (ii) 漢語では連濁が生じない (iii) 複合動詞では連濁が生じない，という 3 つの規則をトップダウンに与え，これらの規則に該当しないものは連濁が生じるとした場合の正解数は 8,583 語，正解率は 85.4 % であった。

[12] 全て連濁するとした場合の成績。

| 利用した素性 | 正解数 | 正解率 (%) |
|---|---|---|
| ベースライン[12] | 2,584 | 50.1 |
| 品詞 ($f_{19}$〜$f_{21}$) | 3,627 | 79.3 |
| 後部要素の音素 ($f_{11}$〜$f_{18}$) | 4,289 | 83.3 |
| 後部要素 ($f_2$) | 4,219 | 81.9 |
| $f_2 + f_{11}$〜$f_{18}$ | 4,398 | 85.4 |
| $f_2 + f_{19}$〜$f_{21}$ | 4,404 | 85.5 |
| $f_2 + f_{11}$〜$f_{18} + f_{19}$〜$f_{21}$ | 4,480 | 87.0 |
| 全素性 | 4,388 | 85.2 |

表 6　素性別の成績 (後部要素が和語のもの)

次に，後部要素を共有する語をもつ 4,710 語について，Ohno (2000) の示した前部要素の類似性が効果をもっているかどうかを見る[13]。

結果は表 7 のようになり，前部要素の品詞情報は正解率を有意に向上させた ($\chi^2 = 8.18, p < .01$) が，前部要素の音素・モーラ数など音韻的特徴による有意な成績改善効果は見られなかった。

| 利用した素性 | 正解数 | 正解率 (%) |
|---|---|---|
| ベースライン[14] | 2,421 | 51.4 |
| 後部要素 ($f_2$) のみ | 4,056 | 86.1 |
| + 前部要素音素 ($f_3$〜$f_{11}$) | 4,026 | 85.5 |
| + 前部要素モーラ数 ($f_{24}$) | 4,085 | 86.7 |
| + 前部要素の品詞 ($f_{19}$) | 4,150 | 88.1 |
| $+f_3$〜$f_{11} + f_{19}$ | 4,038 | 85.7 |
| $+f_{19} + f_{24}$ | 4,175 | 88.6 |

表 7　素性別の成績 (後部要素が和語で，後部要素の共通する語をもつもの)

### 4.3　アナロジーはライマンの法則を維持できるか

ここではアナロジーモデルが誤った予測をしたケースを観察し，モデルの潜在的な問題点を指摘する。

アナロジーモデル (後部要素を和語に限定，全素性を利用) は 56 語についてライマンの法則に違反した予測を行った。誤りの内容を見ると，人間の行動のモデルとしては極めて不自然であることがわかる。例えば「線香花火」という語は，「人情話」と多くの箇所で音素が一致し，また品詞・語種・モーラ数が完全一致したため，これをアナロジーのベースとして選択し，「線香花火」に連濁が生じる (「せんこうばなび」) という誤った予測が行われた。

これは部分的には，本発表で用いた特定のモデルの問題と考えられるが[15]，Albright & Hayes (2003) は，非常に類似度の高い単一の語例の存在に影響されやすいというアナロジーモデルのふるまいが人間の行動を反映していないことを示しており，ここにはアナロジーモデルの本質的な問題が現れている可能性もある[16]。

## 5　おわりに

本発表では形式的アナロジーモデルを日本語の連濁に適用し，その有効性を確かめた。

本発表はアナロジー以外の学習アルゴリズムと比較を行ったわけではなく，特にアナロジーモデルの優位性を示したとは言えない。しかし，アナロジーによる説明と規則による説明の経験的な差を出すためには双方の十分な形式化が必要であり (Albright & Hayes 2003)，本発表のモデルはその一つの方法を示したと考える。

ただし，本発表は既存語の分析にとどまっている。アナロジーが人間の行動モデルとして適切かどうかを検討するには，心理実験と組み合わせて，未知語に対する話者のふるまいをモデルが予測するかどうかを調べることが不可欠である。今後の課題としたい。

[13] Ohno は音韻的類似だけでなく意味的類似も同様の効果をもつとしているが，本発表では意味的特徴を素性として立てることはできていないため，音韻的特徴についてのみ議論する。

[14] 全て連濁するとした場合の成績。

[15] 例えば，ここでのモデルは濁音どうしの類似度が高いという情報を利用できない。実際，後部要素の子音の類似度の計算を MVDM (3.3 節) に切り替えることで，ライマンの法則の違反は 42 に減少した。また，母音の一致を考慮しないようにすることで違反は 5 に減少した。本発表で素性の重み付けに用いた Gain Ratio は，この母音と子音の重要性の違いを検知できていない。

[16] なお，Vance (1980) は，ライマンの法則の心理的実在性が弱いにもかかわらず通時的にパタンが維持されているのは，威信的な一部の話者がライマンの法則を内面化しているためであると示唆している。

## 参考文献


Albright, A. & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, **90**, 119–161.

天野 成昭・近藤 公久 (1999).『NTT データベースシリーズ 日本語の語彙特性』. 東京: 三省堂.

Baayen, R. H. (2003). Probabilistic approaches to morphology. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic Linguistics*, pp. 229–287. Cambridge: MIT Press.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2005). *Memory-Based Language Processing*. Cambridge: Cambridge University Press.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2007). TiMBL: Tilburg Memory-Based Learner Version 6.0 Reference Guide. *ILK Technical Report — ILK 07-03*.

Ito, J. & Mester, A. (2003). *Japanese Morphophonemics: Markedness and Word Structure*. Cambridge: MIT Press.

Krott, A., Baayen, R., & Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics*, **39** (1), 51–93.

Krott, A., Schreuder, R., & Baayen, R. (2002). Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds. In R. Skousen, D. Lonsdale, & D. B. Parkinson (Eds.), *Analogical Modeling: An Examplar-Based Approach to Language*, pp. 181–206. Amsterdam: John Benjamins.

Ohno, K. (2000). The lexical nature of *rendaku* in Japanese. In M. Nakayama & C. J. J. Quinn (Eds.), *Japanese/Korean Linguistics*, Vol. 9, pp. 151–164. Stanford: CSLI Publications.

Otsu, Y. (1980). Some aspects of *rendaku* in Japanese and related problems. In Y. Otsu & A. Farmer (Eds.), *MIT Working Papers in Linguistics*, Vol. 2, pp. 207–227. Cambridge: MIT, Department of Linguistics and Philosophy, MITWPL.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

佐藤 大和 (1989).「複合語におけるアクセント規則と連濁規則」. 杉藤美代子（編）,『日本語の音声・音韻 (上)』, pp. 233–265. 明治書院.

高野 京子 (2006).「日本語の動詞由来複合語におけるアクセントと連濁について」.『日本言語学会 第 133 回大会 予稿集』, pp. 228–233. 日本言語学会.

van de Weijer, J. (2003). East meets west: rendaku voicing in Japanese and linking segments in Dutch compounds. In T. Honma, M. Okazaki, T. Tabata, & S. Tanaka (Eds.), *A New Century of Phonology and Phonological Theory: a festschrift for professor Shosuke Haraguchi on the occasion of his sixtieth birthday*, pp. 268–274. Kaitakusha.

Vance, T. J. (1980). The psychological status of a constraint on Japanese consonant alternation. *Linguistics*, **18**, 245–267.

Vance, T. J. (1987). *An Introduction to Japanese Phonology*. Albany: State University of New York Press.