B-6

共通日本語アクセントは二型アクセントか

情報理論からの観点

竹村亜紀子　神戸大学人文学研究科
Thomas Pellard　CNRS-EHESS-Inalco, CRLAO

要旨

本研究では共通日本語のアクセント型がどれぐらい予測しにくいかを計量的に調査した．アクセント型の分布にモーラ数や語種（和語・漢語・外来語）等によって偏りがあることに着目し，Kubozono (2006; 2008) が推測するように共通日本語アクセントが二型であるといえるかどうかも検証した．コーパスから頻度の高い名詞を 5000 語集め，アクセント型の分布の偏りを情報量という一つの数値で表し，情報量からアクセント型の有効数を計算した．また，条件付き情報量による有効数を使い，語種・モーラ数・音節構造の情報とその組み合わせがアクセント型の予測にどの程度役立つかを測った．その結果次のような点が明らかになった．

1. モーラ数と音節構造という情報を考慮すると，和語 ≫ 漢語 ≃ 外来語の形で表されるようにアクセント型が予測しにくい（和語が一番予測しにくい）．
2. 語種・モーラ数・音節構造のすべての情報がある場合のみ，アクセント型の有効数が 2 以下となり，共通日本語アクセントが二型アクセント体系に相当するといえる．
3. ただし，アクセント型の有効数が 2 以下となるのは漢語と外来語のみで，和語は二型とはいえない．

1　はじめに

共通日本語は語彙ごとにアクセント型（アクセント核の有無とその位置）が決まっており，予測できない自由アクセント体系である．しかし，アクセント型の分布には語種（和語・漢語・外来語），長さ（モーラ数），音節構造等によって偏りが見られ，語末から 3 モーラ目にアクセント核がくるデフォルトのアクセント型（–3 型）が存在する（Martin 1952; 柴田 1994; 佐藤 1993; 栗林 1996; Kubozono 2006; 2008; Ito & Mester 2016）．つまり，アクセント型は全く予測できないという訳ではないのだが，どの程度予測可能かが未だに明らかにされていない．

Kubozono (2006; 2008) では 3 モーラ名詞（7937 語）のアクセント型の頻度を語種別に調べた結果，語種によらず平板型（0）が大多数であるが，起伏型をもつ漢語と和語では –3 型が多いことが明らかになっている（図 1）．つまり，例外はあるものの，共通日本語アクセントは二型アクセント体系（平板型と –3 型の二つのアクセント型）に近いとされている．しかし，二型アクセント体系に完全に収まらないので，二型アクセントにどれぐらい近いかという問題が残されている．

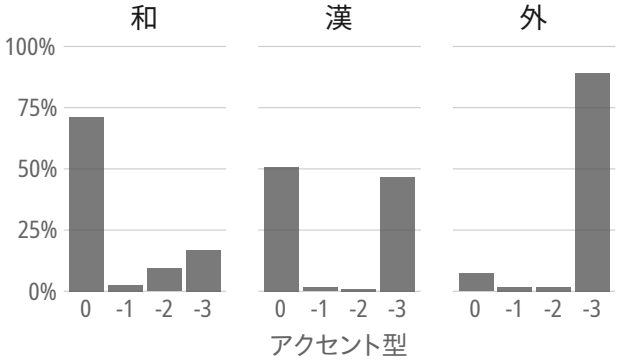

図 1　3 モーラ名詞（7937 語）における語種別のアクセント型の分布（Kubozono 2006; 2008 をもとに作成）

この問題を解決し Kubozono (2006; 2008) の推測を検証するために計量的な観点が必要であるが，それぞれのアクセント型の頻度を計算しただけではその分布全体の偏りの程度が分からず，語種の間の比較も難しい．そこで本研究では，この問題を克服するために情報量という概念を用いてアクセント体系全体及びそれぞれの語種におけるアクセント型の不確実性を計算し，区別されているアクセント型の有効数という概念を導入する．また，語彙リスト（辞書）におけるアクセント型の頻度だけではなく，コーパスにおける頻度にも着目する．

2　データ

この研究では BCCWJ （Maekawa et al. 2013）の語彙表（頻度リスト）から普通名詞（長単位）のリストを取り出し，UniDic （伝 et al. 2007）と結合させてアクセント等の情報を得た．[1)] UniDic には短単位しか記載されていないので，長単位でも出現している短単位のリストが得られた．そうすることによって，多くの合成語はリストに含まれなかった．アクセント型が強く偏っている普通名詞以外の語，明らかな派生語・複合語，7 モーラ以上の語彙の他に標準的でない語形（変種）や均質的な語種を成さない頭文字語・混合語を省いた最頻名詞の 5000 語（異なり語数），延べ語数では 10 767 958 語を抽出した．[2)] アクセント型が複数記載されている場合，最初のもののみを利用した．また，アクセント情報を語末から数えた位置に換算し，モーラ数と音節構造（軽･重音節）[3)] の情報を追加した．

図 2 ・3 ・4 はそれぞれ本研究のデータにおける語種・モーラ数・アクセント型の頻度（述べ語数及び異なり語数）を表している．これらを見ると，漢語は延べ語数でも異なり語数でも出現する割合が高いことがわかる．一方，和語は異なり語数では少ない（つまりタイプは少ない）が延べ語数では多いという特徴がある．そして，外来語は延べ語数でも異なり語数でも出現する割合が相対的に低い．また，異なり語数では 4 モーラ語が多く，延べ語数では 2 モーラ語が多い．さらに，異なり語数でも延べ語数でも平板型（0）が圧倒的に多い． 図 5 はモーラ数と語種別のアクセント型の頻度をまとめて示している．

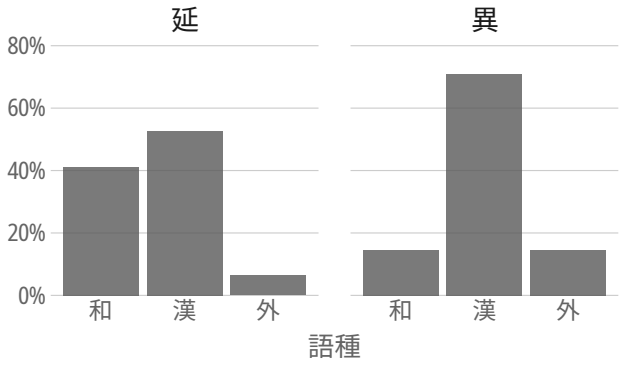

図 2　本研究のデータにおける語種の頻度

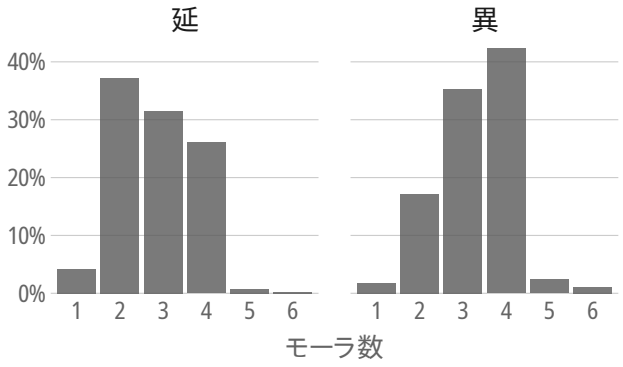

図 3　本研究のデータにおけるモーラ数の頻度

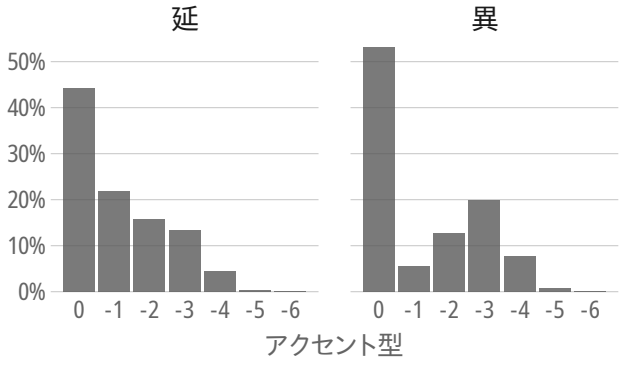

図 4　本研究のデータにおけるアクセント型の頻度

1)　データの抽出や結合及びその分析には R （v. 4.1.0; R Core Team 2021）の tidyverse （v. 1.3.1; Wickham et al. 2019）を使用した．

2)　アクセント型が規則で決まっているものをあえて排除した場合，アクセント型がどれぐらい予測しにくいかに着目した．

3)　語形を L （軽）･H （重）･S （超重）の羅列に変換した．

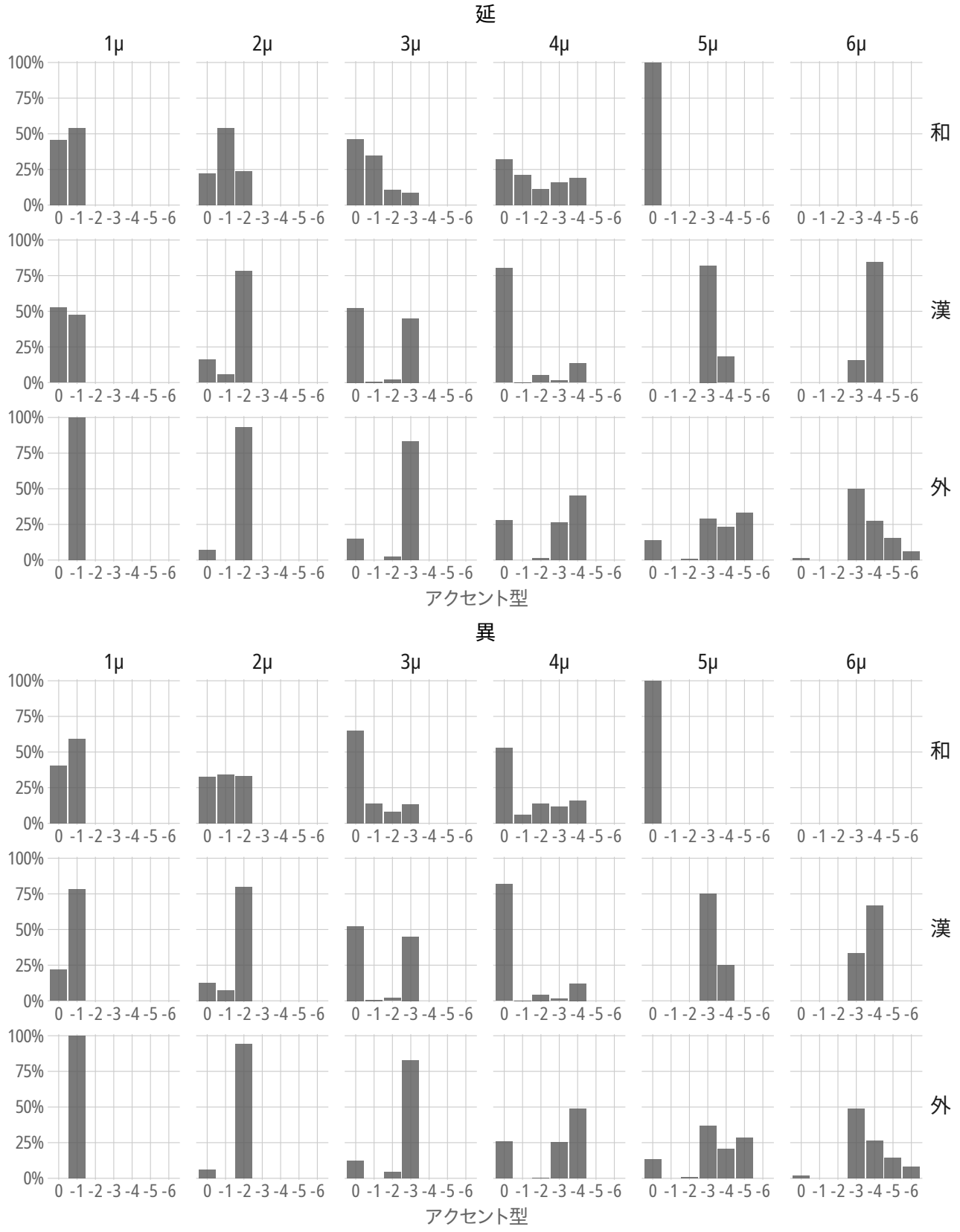

図 5　本研究のデータにおけるモーラ数（μ）別・語種別のアクセント型の頻度

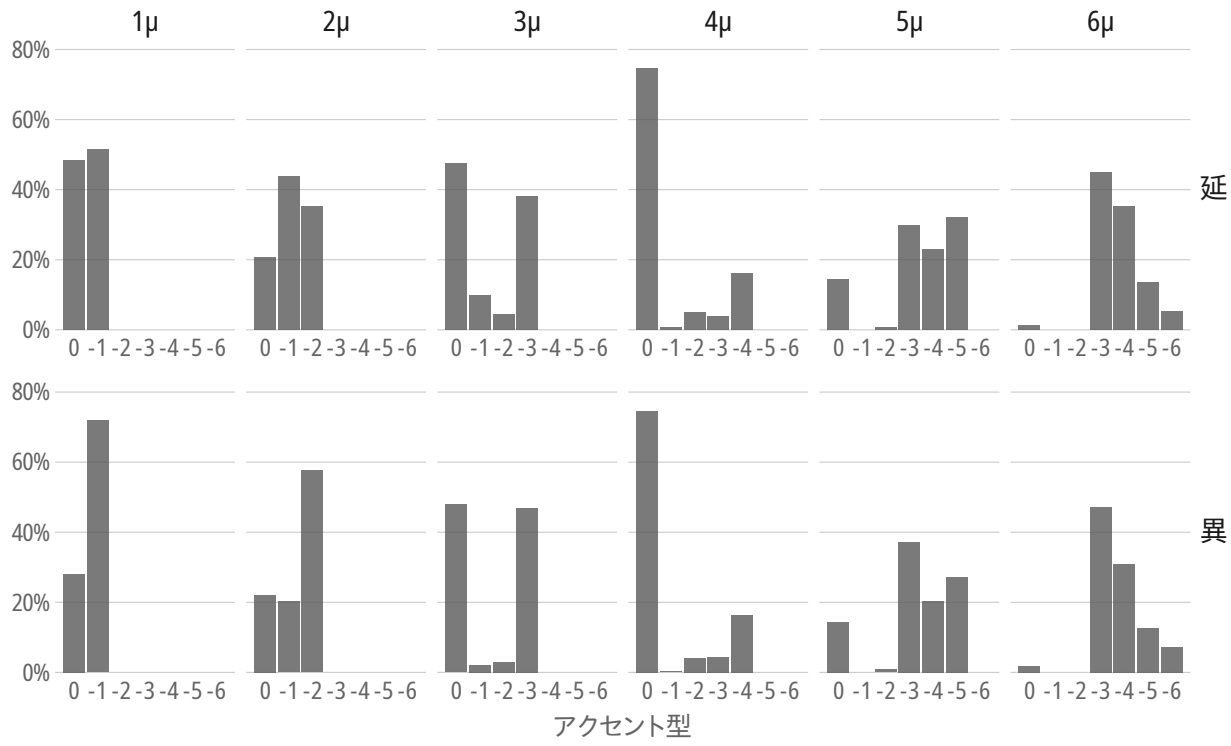

図 6　本研究のデータにおけるモーラ数別のアクセント型の頻度

3　方法

3.1　情報量

本研究ではアクセント体系全体及びそれぞれの語種におけるアクセント型の偏りを数値化するために情報理論（Cover & Thomas 2006 等）を用いる．

まず，アクセント型の**情報量**（entropy）を計算するが，情報量とはある確率変数（ここではアクセント型）の結果が平均的にどのぐらい予測しにくいか，その分布がどのぐらい偏っているかを表す尺度である（単位は bit）．それぞれの結果が同じ頻度（確率）で起こる場合，結果がもっとも予測しにくく情報量が最大となる．可能な結果の数が n であれば，情報量の最大値が $\log_2 n$ となる．一方，結果の分布に偏りが強ければ強いほど結果が予測しやすくなるので情報量が下がる．確率変数 X の情報量 $H(X)$ はそれぞれの結果の確率（$P(x)$）から（1）の式のように求められる．

$$(1)\quad H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

しかし，アクセント型の頻度をそのまま使って情報量を計算すると次の問題が起こる．まず語の長さ（モーラ数）によって可能なアクセント型の数が大きく異なり，[4] アクセント型の頻度も大きく異なる結果（図 6），情報量の値がモーラ数によって大きく左右されるという問題がある（図 7）．さらに，モーラ数が語種によって大きく異なるという問題もある（図 8）．

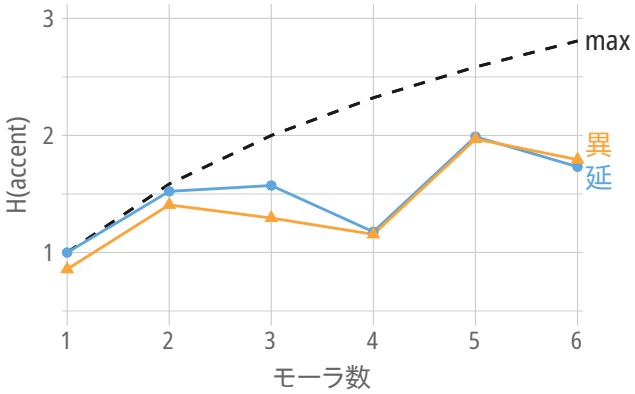

図 7　モーラ数別のアクセント型の情報量

4）n モーラに対して可能なアクセント型が最大 $n+1$ 個ある．

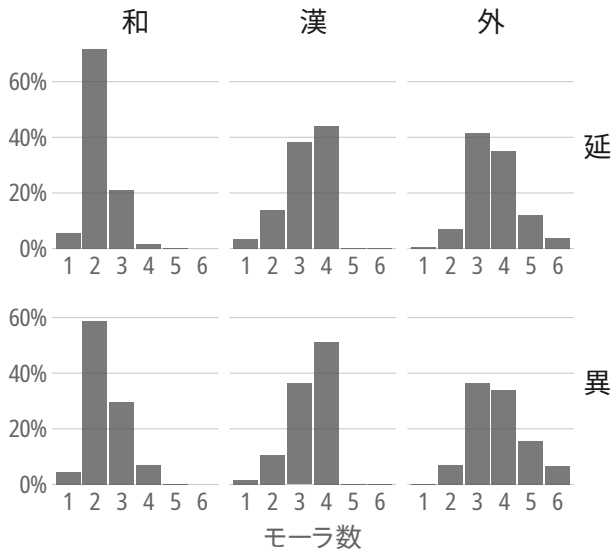

図 8　本研究のデータにおける語種別のモーラ数の頻度

3.2　条件付き情報量

このようにいくつかの情報が複雑に絡み合っている事象を扱う場合は**条件付き情報量**がより妥当である．条件付き情報量 $H(X \mid Y)$ は，単純な情報量 $H(X)$ と異なり当該の変数（X）の他に，別の変数（Y）の情報を考慮に入れて計算する．例えば，単語の長さ（モーラ数）を知った時にアクセント型がどれぐらい予測しにくいかを計算する．具体的に条件付き情報量 $H(X \mid Y)$ は X と Y のそれぞれ組み合わせの確率（例えば 2 モーラ平板，2 モーラ –1，2 モーラ –2…）を基に計算される結合情報量 $H(X,Y)$（2）から Y の情報量 $H(Y)$ を引いて求められる（3 ・図 9).

$$(2)\quad H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(x,y)$$

$$(3)\quad H(X \mid Y) = H(X,Y) - H(Y)$$

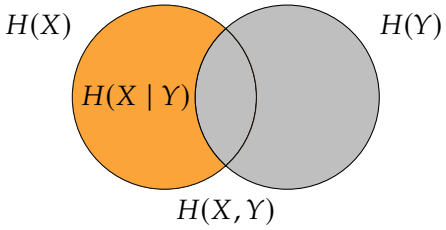

図 9　条件付き情報量

モーラ数によってアクセント型の分布が偏っているのは確かだが，条件付き情報量を使うことによってこの両現象が関連している問題が解消される．本研究ではモーラ数の他に音節構造と語種の情報とそれらの組み合わせの条件付き情報量を計算し，どの組み合わせがどの程度アクセント型の予測に役立つかを測った．

3.3　有効数

情報量という尺度は馴染みがなく，直感的に分かりにくいというデメリットがある．このような場合，代わりに**有効数**（effective number）を使うのが便利である．有効数とは頻度が少ないカテゴリーより頻度の高いカテゴリーを重視する多様度指数で，例えば生物の種の多様性を測るのに使われている（Magurran 2004; Roswell et al. 2021）．ここでは有効数を周辺的・例外的なアクセント型を重要視せずアクセント型の数がいくつあるかを示すために用い，Kubozono (2006; 2008) の推測を直接的に検証できるようにする．有効数は具体的に（4）のように情報量から求められる．モーラ数や音節構造等から予測しやすいアクセント型に左右されない，条件付き情報量に基づく有効数も同じように計算できる．

$$(4)\quad D(X) = 2^{H(X)}$$

4　結果

モーラ数，音節構造，語種の条件を加えてアクセント型の有効数を計算した結果を図 10 に提示する．まず，異なり語数と延べ語数のどちらで計算しても，大きな差が見られない．モーラ数の情報を足しただけで有効数が大きく下がり，2 型以上 3 型未満（延：2.69，異：2.41）となる．さらに音節構造と語種の条件を比べると，延べ語数の場合のみ僅かな差が見られる．モーラ数，音節構造，語種の 3 つの条件を全て加えると有効数が延べ語数では 2.28，異なり語数では 2 以下となる．

共通語でアクセント核が付与されるのはモーラなのか音節なのかという議論が長らく続いている

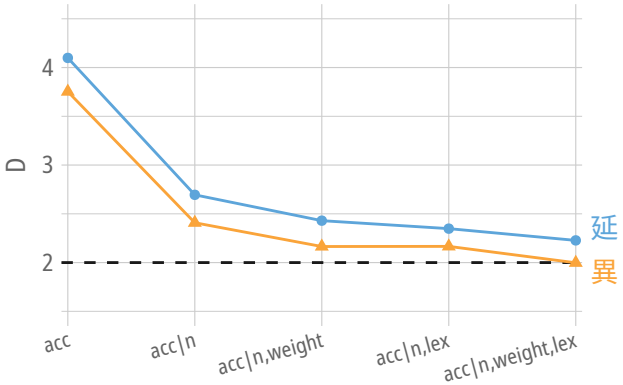

図 10　アクセント型の有効数（acc = アクセント型，n = モーラ数，weight = 音節構造，lex = 語種）

ため（McCawley 1968; 上野 2003 等），モーラではなく音節でアクセント核の位置を数えた場合の有効数の計算も行った（図 11）．しかし両者の間にあまり差が見られないので，これ以降もモーラでアクセント核の位置を数えることにした．

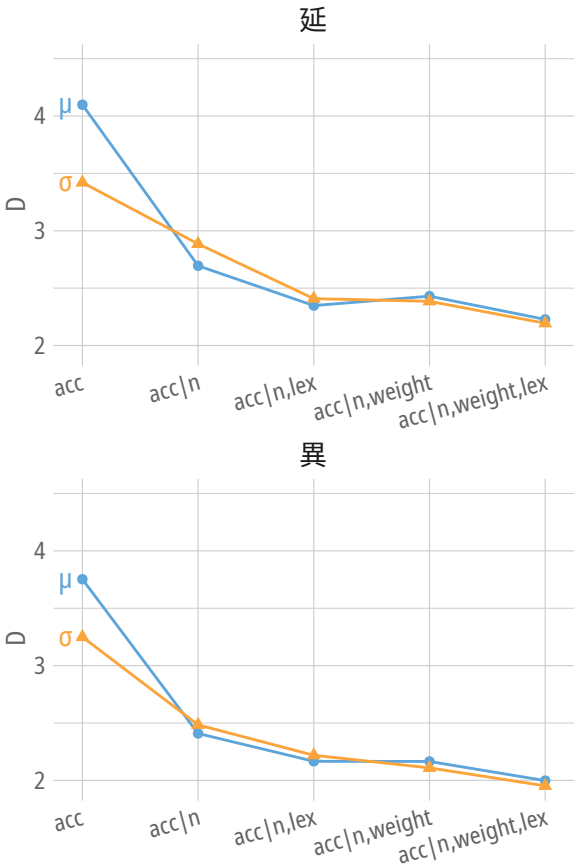

図 11　モーラ（μ）と音節（σ）に基づくアクセント型の有効数

図 12 では語種別にアクセント型の有効数を計算した．ここでも延べ語数でも異なり語数でも結果はだいたい同じである．漢語と外来語はモーラ数と音節構造の条件を加えるとそのアクセント型の有効数が 2 以下となるが，和語は 2.7（延）または 2.81（異）で 2 以下にはならない．

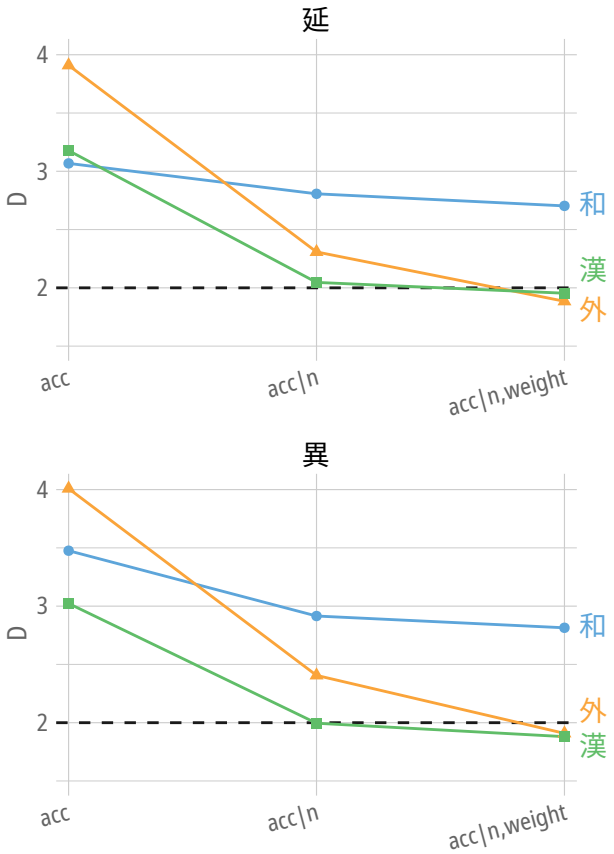

図 12　語種別のアクセント型の条件付き有効数

5　考察と結論

本研究では共通日本語のアクセント型がどの程度予測しにくいかを情報理論の観点から計量的に調査した．アクセント型の分布の偏りとその予測しにくさを情報量で表し，条件付き情報量でモーラ数・語種・音節構造という情報がアクセント型の予測にどの程度役立つかを測った．さらに，有効数という指標を用いることで，有効数＝主なアクセントの型の数という比較しやすい形で提示することが可能であることも示した．

本研究で明らかになったのは以下の点である．

1. 情報量及び有効数を異なり語数と延べ語数の頻度で計算しても，両者の間に大きな違いが見られない．

2. アクセント核の位置をモーラで計算しても音節で計算しても大きな差が見られない．
3. アクセント型の予測にはモーラ数が大きく貢献し，語種と音節構造の貢献が同じ程度である．
4. モーラ数と音節構造の条件を加えると和語 ≫ 漢語 ≃ 外来語の順にアクセント型が予測しにくい（和語が一番予測しにくい）．
5. 共通日本語アクセントが二型アクセントに相当するという Kubozono (2006; 2008) の推測はモーラ数・語種・音節構造のすべての情報を考慮した場合のみ妥当だといえる
6. 語種別に見ると漢語と外来語は二型といえるが，和語はそうではない．

和語が漢語と外来語とは異なる振る舞いを示しており，アクセント型の有効数が常に 2 以上でより予測しにくいことが分かった．その理由は，現代の和語のアクセント型は歴史の産物であり，漢語や外来語と違い，生産的な規則の結果ではないからであると思われる．

有効数の観点から共通日本語アクセントが二型アクセントに相当すると言っても，それがアクセント型には選択が平均二つであるということで，有効な二つのアクセント型がモーラ数・語種・音節構造を通して同一であるとは限らないことに注意が必要である．

謝辞

本研究は科研費・研究基盤（C）「日本語の音韻の機能負担量に関する計量的研究（代表: 竹村亜紀子）」（#19K00644）の支援を受けたものです．

参照文献

Cover, Thomas M. & Thomas, Joy A. (2006) *Elements of information theory*. 2nd edn. Hoboken: Wiley.

伝康晴 àtext ・小木曽智信 àtext ・小椋秀樹 àtext ・山田篤 àtext ・峯松信明 àtext ・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』 22: 101–122.

Ito, Junko & Mester, Armin (2016) Unaccentedness in Japanese. *Linguistic Inquiry* 47(3): 471–526. https://doi.org/10.1162/LING_a_00219.

Kubozono, Haruo (2006) Where does loanword prosody come from? A case study of Japanese loanword accent. *Lingua* 116(7): 1140–1170. https://doi.org/10.1016/j.lingua.2005.06.010.

Kubozono, Haruo (2008) Japanese accent. In Miyagawa, Shigeru & Saito, Mamoru (eds.), *The Oxford handbook of Japanese linguistics*, 165–191. New York: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195307344.013.0007.

栗林均（1996）「現代日本語のアクセントの型の分布：『電子ブック版 大辞林』を資料として」『研究紀要』 51: 1–28.

Maekawa, Kikuo & Yamazaki, Makoto & Ogiso, Toshinobu & Maruyama, Takehiko & Ogura, Hideki & Kashino, Wakako & Koiso, Hanae & Yamaguchi, Masaya & Tanaka, Makiro & Den, Yasuharu (2013) Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation* 48(2): 345–371. https://doi.org/10.1007/s10579-013-9261-0.

Magurran, Anne E. (2004) *Measuring biological diversity*. Malden: Blackwell.

Martin, Samuel E. (1952) *Morphophonemics of Standard Colloquial Japanese* (*Language* Dissertation 27). Baltimore: Linguistic Society of America. https://doi.org/10.2307/522176.

McCawley, James D. (1968) *The phonological component of a grammar of Japanese*. The Hague: Mouton.

R Core Team (2021) *R: A language and environment for statistical computing*. Computer software. https://www.R-project.org.

Roswell, Michael & Dushoff, Jonathan & Winfree, Rachael (2021) A conceptual guide to measuring species diversity. *Oikos* 130(3): 321–338. https://doi.org/10.1111/oik.07202.

佐藤大和（1993）「共通語アクセントの成因分析」『日本音響学会誌』 49(11): 775–784. https://doi.org/10.20697/jasj.49.11_775.

柴田武（1994）「外来語におけるアクセント核の位置」佐藤喜代治(編)『現代語・方言の研究』（国語論究 4）418–388．東京：明治書院.

上野善道（2003）「アクセントの体系と仕組み」上野善道(編)『朝倉日本語講座 3：音韻』61–84．東京：朝倉書店.

Wickham, Hadley (2019) Welcome to the tidyverse. *Journal of Open Source Software* 4(43): 1686. https://doi.org/10.21105/joss.01686.