

# 英単語の発音と綴りの, HDP を使った教師なしパターン抽出

## Unsupervised extraction of sound and spelling patterns for English words using HDP

黒田 航 (杏林大学医学部)

### 1 はじめに<sup>1)</sup>

#### 1.1 英語習得の効率化に更に何が必要か？

多くの日本人は, 中学校で3年, 高校で3年, 大学で少なくとも2年の合計8年間, 英語を学ぶ。なのに, 日本人の平均的な英語の習熟度は決して高くない(自分の知っている範囲で言えば, 研究者であっても, 英語が得意でないと言う人は多い)。なぜか? 原因は, 認知科学/学習科学の観点から考えると次だろうか?

- (1) a. 英語が必修科目として否応なく学ぶのに,
- b. 学習指導に, 英語が学習者の母語である日本語と大きく違っている事から来る不利が反映されていない。

認知科学/学習科学が問題(1a)の解決のためにできる事はないが, 問題(1b)の解決のためにできる事はそれなりにあると筆者は考える。本研究の究極の目標は, 問題(1b)の解決に貢献する事である。

現実的な目標として, 本研究は次を定める。

- (2) 英単語の構成パターンを, 発音と綴りの両方の側面から抽出し, 広く参照可能なデータとする。

このような目標を定めている理由は, 日々の自分の英語教員としての活動, 自分自身を観察対象者とした実験を通じて, 次の事を日々実感している<sup>2)</sup>からである:

- (3) (日本語を母語とする) 英語学習者が抱える最大の困難は, 英単語がなかなか覚えられない事である<sup>3)</sup>。

端的に言うと, 英語の語彙要素は明らかに日本語の語彙要素と似ていない<sup>4)</sup>。これが日本人が英語の修得に苦

<sup>1)</sup> 発表応募原稿に対する二名の匿名査読者からのコメントに感謝する。一名のそれは特に改訂に役立った。

<sup>2)</sup> 今のところ実証的な根拠を示す事はできない。

<sup>3)</sup> これは見方を変えると, 異国語修得の際の文法規則の重要性は(理論言語学の主張とは裏腹に) 語彙要素(e.g., 単語)の獲得の重要性には及ばないという事を意味する。

<sup>4)</sup> 異国語の修得で語彙要素が似ていると学習がどれ程楽になるか, 筆者は日本語と系統の異なる言語を10種類以上学んだ後に最近になって中国語を学び始め, それを実感した。中国語の音韻は日本語に全然似ていないが, 統語的には語順がほんの少し違う日本語と言って良い。

労する根本的な理由だろうと推測する<sup>5)</sup>。

語彙要素の獲得が困難である時, それには二つの側面がある。第一に, 学習対象の言語と学習者の母語が語彙要素を共有する程度ある。これには発音と綴りの二つの面があるが, その区別をここでは問題にしない。英語と語彙要素を共有する度合いが高い言語は印欧語派の言語に限られる。これは, これに該当する言語を母語とする学習者は相対的に有利である事を意味する。

第二に, 英語では綴りと発音の対応が(並外れて)悪い事である[1, 2]<sup>6)7)</sup>。これは英語を異国語として学ぶ学習者の母語が何語であっても成立するが, 母語の発音体系が英語のそれに似ている程度に応じて緩和される。とは言え, 発音体系が似ている言語は, 系統も近く, 語彙要素を共有する傾向があるので, 印欧語を母語する話者が相対的に英語学習で有利なのは, 明らかである。このように不利な状況で(3)の制約を緩和するには何ができるか? を考える必要がある。

これは今までにない新しい単語帳を作れば解決する問題ではない。どんなに良い単語帳でも, それは根本的には(3)の解決にはならない。本研究ではそれに代わり語彙獲得に必要なだが, 単語帳に欠落している情報を補う方略を考える。具体的に言うと, 英語の単語の発音と綴りのパターンを抽出し, それを学習者に一緒に辞書として提示する方法を試す。それが(2)に示した目標である。

#### 1.2 パターン抽出の方法

抽出したパターンが有効かは検討は独立に行う事にして, パターン抽出の要件はパターン抽出は大規模データから, 実質的に自動で行えるものでなければならない。現実問題として, そうでない条件で抽出されたパターンに実用的は期待できない。この要件を満足する手法は自

<sup>5)</sup> 日本人が英語の修得で苦労している根本的な理由は, 第二言語習得の語彙障壁(lexical barriers to second language acquisition)に阻まれているからである。

<sup>6)</sup> 黒田[3]は英語の綴りと発音の乖離の定量評価を試みた萌芽的研究である。

<sup>7)</sup> 筆者は長く英語より発音と綴りの対応が悪い言語はないと思っていたが, アイルランド語を学び始めて状況が英語より酷いと知った。



図1と図2を見比べると、次の事がわかる。skippy  $n$ -gram を使った解析は抽象的なパターンを抽出し、かつパターン階層化を表現している。外側にある、半径の大きなパターン (= term) は抽象度が高い。

図1に関して言う、明らかに方向性と次元性がある。数の大きな (= 優先順位の低い) トピックは、一定に凝集している。その一方、トピック1とトピック2は別の方向を向いている。これは数の小さい (= 優先順位の高い) トピックは、基本ベクトルのような働きをしている事を示唆する。なお、これは2D可視化だが、多次元可視化では、トピック1, 2, ..., 10 ぐらいまでは基本ベクトルのように振る舞っているように見える。図2には、方向性も次元性も認めにくい。この解析でトピックとなるのは音節的な要素である。

## 2.2 綴り文字の HDP の解析結果

図3に term を綴り文字の不連続 5-gram とし、topic 数の上限を 90 に設定した HDP の結果の pyLDAvis を使った可視化 (topic 9 を選択した状態) を示す。

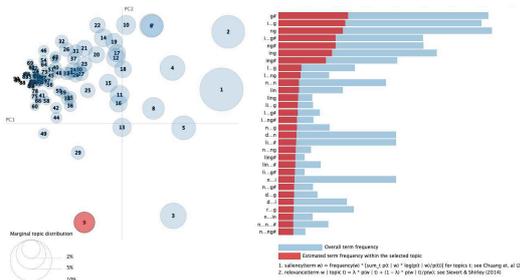


図3: term: skippy 5-gram of letters + #

図4に term を文字の連続 5gram とし、topic 数の上限を 90 に設定した HDP の結果の pyLDAvis を使った可視化 (topic 4 を選択した状態) を示す。

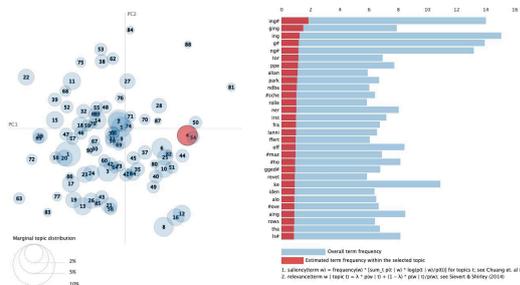


図4: term: (continuous) 5-gram of letters + #

### 2.2.1 綴り文字の HDP の結果に基づく考察

図3と図4を見比べると、skippy  $n$ -gram を使った解析は抽象的なパターンを抽出し、かつパターン階層化を表現している事がわかる。これは IPA 記号列を term とした解析結果と同じである。term の長さに固定化の傾向が認められないのも、同じである。

ただ、発音パターンの場合と違い、半径の大きな外側にあるトピックがどんなパターンを抽出しているのかは、解釈が難しい。特に、図4のトピックが何を表わしているかは、解釈がしづらい。それでも、様々なパラメータで得た結果を比較して、次の事は言える。

- (4) a. HDP/LDA を使ったパターン抽出は、i) 連続  $n$ -gram を使った場合には (生起位置に関係なく) 形態素 (~ 語彙素) を抽出する傾向があり、逆に ii) 不連続  $n$ -gram を使った場合には、配置に関係なく # を含む抽象的なパターンを抽出する傾向がある<sup>11)</sup>。
- b. 図3の広がりの方角性は、語形と品詞の相関に対応しているらしい。具体的に言うと、方向性は名詞的な語末、動詞的な語末のような違いに対応しているらしい。

(4a) から判断すると、文字の連続  $n$ -gram と不連続  $n$ -gram を term に使った HDP からは、相補的な結果が得られていると考えて良い。これらをうまく組み合わせると有用な学習資源を構築できると期待できる。

(4b) の想定に着想を得て、英語の語彙を品詞別に下位分類し、それぞれに上記と同じ条件で HDP を適用した結果を得ている。品詞ごとに典型的な終わり方と始まり方があり、解析結果はこれを捉えている。紙面の都合で本稿では詳細を述べられないが、形容詞に関してのみ、図5に結果 (topic 3 を選択した状態) を示す<sup>12)</sup>。

## 3 議論

結果の考察とは別に、次の理論的な点について、私見を述べて置きたい。LDA/HDP はトピックモデル (topic models) [7, 8] である。トピック (topic) は通例、意味的なものだと考えられるが、単語の発音や綴りにトピックがあるという想定は妥当なのか？

<sup>11)</sup> これは英語でも語の始まり方と終わり方にパターンがある事実に対応している。

<sup>12)</sup> open-dict-ipa のデータは語形のみを提供し、品詞情報は持っていない。語形を noun, verb, adjective, adverb への下位分類する処理は WordNet (3.0) の情報を参照して行った。

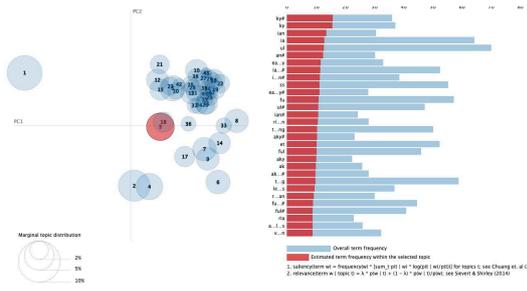


図 5: term: skippy 4-gram of letters + # (adj only)

トピックモデルは文書分類の手法として発展して来た。トピックが何らかの意味に対応しているという想定は、この背景から理解すべきである。トピックモデルがやっている事は、全体となる集合(例えば文書)とそれらの構成部分(例えば単語)との共起を説明する潜在因子の推定である。見つかる因子は、概念的な意味と言うより、より抽象的に潜在生成源と考えるべきである。研究[4]で得られている結果は、そのように解釈できる。

この再解釈が必要となる証拠を追加する。図6に示したのは、単語  $w$  の表記  $w.sp$  と発音  $w.sn$  の対 (e.g., people/pipl, thing/θing, ...) を documents とし,  $w.sp$  の skippy 3gram と  $w.sn$  skippy 3gram の対を term とし<sup>13)</sup>, topic 数の上限を 90 に設定した HDP の結果の pyLDAvis を使った可視化 (topic 23 を選択した状態) である。これは綴り “qui...(t)” と発音 /kwɪ...t/ の対応を分散的に抽出している。この結果は、綴りと発音の有意な対応を自動抽出できる事を示唆している<sup>14)</sup>

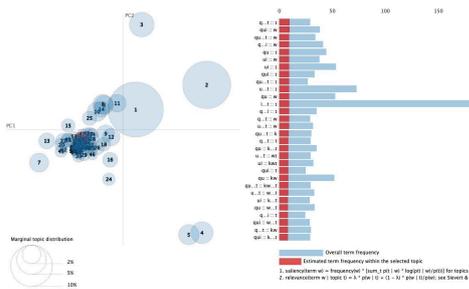


図 6: term: pair of skippy 3-grams for  $w.sp$  and  $w.sn$

このような高次の対応関係の抽出が LDA/HDP で可能である事は、トピックが部分の共起関係を説明する因子に対応しているという再解釈の妥当性を裏づける。

<sup>13)</sup>  $w.sp$  の  $n$ -gram と  $w.sn$  の  $n$ -gram との対は総当たりで構築している。計算量が相当に多いため、HDP では処理に相当の時間がかかるが、LDA ではそれ程でもない。

<sup>14)</sup> 研究 [3] は人手による分割を使って半手動で同じような対応関係を抽出している。産物の比較は今後の課題である。

## 4 まとめ

本研究では Latent Dirichlet Allocation (LDA) のパラメータなし版である Hierarchical Dirichlet Process (HDP) を使って英単語の(発音記号か綴りの)一覧から、教師なしで (1) 語の表記の発音のそれぞれの構成パターンと (2) 綴りと発音の対応関係が自動抽出可能であるという結果を得た。現時点では、開発目標の日本人の英語の語彙獲得を支援する学習資源は構築できていないが、その手始めにはなっているように思われる。目標の学習資源を提供するためには、パターンの品質の評価が必要である。

HDP/LDA を使った語構成パターン、綴りと発音の対応関係の自動抽出は、言語を選ばない一般性がある。データさえあれば、どんな言語にも同じ解析手法が適用できる<sup>15)</sup>。これは一般に異国語学習の学習支援データを(半)自動で構築できる可能性を示唆する。

提案手法の明らかな限界は、抽出されたパターンが(分散的であるが故に)冗長である点にある。これはパラメータの設定で対処できる問題ではない。産物の有用性の向上には、後処理で冗長性を削減する必要がある。

## 参考文献

- [1] Jane Chinelo Obasi. Structural irregularities within the English language: Implications for teaching and learning in second language situations. *J. of English as an International Language*, Vol. 13, No. 2.1, pp. 1–14, 2018.
- [2] Arika Okrent. *Highly Irregular: Why “Tough,” “Through,” and “Dough,” Don’t Rhyme and Other Oddities of the English Language*. Oxford University Press, 2021.
- [3] 黒田航. 英単語の綴りと発音のズレの定量的評価. 認知科学会第 40 回大会発表論文集, pp. 644–647, 2023.
- [4] Kow Kuroda. Revealing regularities in spelling and pronunciation among languages using Latent Dirichlet Allocation. In *Proc. of the 30th Annual Meeting of the Natural Language Processing Association*, 2024.
- [5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.
- [6] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581, 2006.
- [7] 岩田具治. トピックモデル. 講談社, 2015.
- [8] 佐藤一誠. トピックモデルによる統計的潜在意味分析. コロナ社, 2015.

<sup>15)</sup> GitHub サイトには提案手法で、アラビア語、アイルランド語(綴りデータのみ)、フランス語、ドイツ語、スペイン語、スワヒリ語の単語の構成パターンを抽出した結果を挙げてある。