英単語の綴りと発音のズレの定量的評価

Quantitative assessment of the spell-sound discrepancy among English words

黒田　航
杏林大学医学部

概 要

CMU Pronouncing Dictionary を使って，単語の綴りと発音の対応が理想的な 1 対 1 対応から外れている程度を，二つの方法で定量評価した．一つは，必須度の高い資源中の 4,290 語にの綴りと発音の対応の定量評価で，もう一つは英語とドイツ語の高頻度語形を同じ条件で比較した．これらから，英語での対応が一対一から大きく外れている事がわかった．

1　英単語はどうして覚えにくいのか？

多くの日本人は英語を学ぶのに苦労する．発音は日本語と大きく違うし，文法も大きく違う．だが，それ以前に，多くの日本人は英単語を覚えるのに苦労する．英語の音が日本語の音に似ていないのは理由の一部だろうが，それ以前に英単語では**綴りと発音の対応の不規則性が尋常でない** (Obasi, 2018)．そのため「英単語はどうして (こんなに) 覚えにくいのか？」と感じている学習者は数多いはずである．その理由を明らかにするために，本研究では英単語の読み (= 発音) と綴りの対応が規則的でない程度を定量的に示す．

1.1　関連研究

英語の綴りの不規則性に関しては，第二言語としての英語 (English as a Second Language: ESL) 教育の文脈で，学習者が見せる綴り間違いの観点からの記述 (Wolf, 2017) が多い．ただ，これらの研究は，英語の単語の発音と綴りの対応の悪さそれ自体を考察，記述してはいない．

英語単語の発音と綴りの対応の悪さへの言及は，学術研究に対象を限定しなければ，それなりに多い (Patterson, n.d.; *Highly irregular*, n.d.) が，それに体系的，かつ網羅的に言及した文書は少ない (例外は (*Highly irregular*, n.d.))．(Okrent, 2021) は学術研究として，この問題を扱っているが，定量化を試みていない．不規則性の程度を，発音と綴りの要素の分布を基に定量的に評価した先行研究は見当たらない．

1.2　定量評価のための設定

定量化のために，単語の学びにくさ/やすさを次のように定義する:

(1) 単語は，(意味上の親しみやすさなどの他条件が同じであれば) 発音と綴りの対応が規則的である程，覚えやすい．

(2) 単語の発音と綴りの対応は，i) 一対一対応に近い程，規則的である．ii) 文脈自由度が高い程 (= 文脈依存度が低い程)，規則的である．

本調査の目的は，英語の綴りと発音の対応では (2) が満足されていない事を定量評価し，その事から (1) が満足されていない事を示す事である．

英語の綴りと発音の対応の悪さは，同じ文字に幾通りかの異なる読みが割り当てられる事で，逸話的に語られる (先に紹介した先行研究の大半がこのような例証である)．だが，このような事例を幾ら列挙しても定量評価は達成されない．そのためには，読みの単位を揃え，網羅的に対応関係を知らなければならない．この目的を実現するために，CMU Pronouncing Dictionary (CMUPD)[1)] を使って，英単語の綴りと発音の対応関係を，単語よりも小さな規模で定量評価し，対応関係の不規則性を定量化し，その結果を評価するのが，本研究の内容である．

論文の構成は次の通り．§2 で CMUPD を使ったデータ構築の説明する．§3 で解析方法と得られた結果を示す．§4 で結論と展望を述べる．

解析に用いたデータと作業に使ったスクリプトは GitHub repository [2)]で公開している．

[1)] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2)] https://github.com/kow-k/English-spell-sound-discrepancy

2 CMUPDを使ったデータ構築

解析では, 英単語4,290語形[3]を対象とした．これを英単語のSpell-Sound対応づけデータ D[4]と呼ぶ．D の語形は次のように幾つかの源から集め, 重複を省いたものである: SketchEngine[5]で獲得した1) EnTenTen20コーパスの上位1k words, 2) EnTenTen20の上位1k lemmas, 3) TEDコーパスの上位1k words，4) The Longman Defining Vocabulary[6] の2,197 lemmas (+10 prefixes, 39 suffixes)，5) General Service List (GSL)[7] の2,284 lemmasに，6) CMUPDの幾つかの要素を追加.

上の4,290語形にCMUPDを使って発音記号を割り当てた．単純な対応づけではなく, 次の処理を経ている.

まず, CMUPDは音素のエンコードにIPA記号を使っていない．代わりにARPABETという2文字エンコーディングを採用している．今はUTF-8で発音記号が処理できる時代なので, 自前のPerlスクリプトでARPABET表記をIPA表記に変換した.

CMUPDが収録しているのは基底/深層形の発音と言うべきもので，実際の発音を収録していない．(3) にamerican, treatmentの発音記述[8]の実例を示す.

(3) i) american, ʌ0mɛ1ɹɪ0kʌ0n
ii) treatment, tɹi1tmʌ0nt

実際の発音 (の近似値) を得るには，(4) のような (母音の弱化を反映した) 記述を導出する必要がある.

(4) i) american, ə0mɛ1ɹɪ0kə0n
ii) treatment, tɹi1tmə0nt

この書き換え処理を自動実行するシステムは見つからなかったので，自前で処理した．その結果が先のSpell-Sound対応づけデータ D のfield 2の値である．具体的には, 表層形を生成するための作業シート W[9] を使った実際の発音になるべく近い発音を自動生成した．W のw.IPA列の値がCMUPDに忠実なIPA表記で，w.IPA.r4列の値が (w.IPA.r0, w.IPA.r1, ...) を経て導出された実際の発音に近いと想定する表記である．なお，w.slashedの値が対応づけデータ D のfield 3の値である[10].

CMPUDの発音記述には，語中の強勢の位置を表すために0 (ground), 1 (primary), 2 (secondary) の記号が使われている．これらは原則として母音の直後に現れる．この特徴を利用すると，音節構造がO(nset) + N(ucleus) + C(oda) だとして，発音をON単位で自動分割可能である (綴り中の，発音記号中の0, 1, 2に対応する位置に何らかの区切り (例えば“/”) を入れると，ON単位での，発音と綴りの対応づけが体系的に獲得できる)．具体的には，次の通り:

(5) i) ə0mɛ1ɹɪ0kə0n, a/me/ri/ca/n ⇒ [ə:a, mɛ:me, ɹɪ:ri, kə:ca, n/n]
ii) tɹi1tmə0nt, trea/tme/nt ⇒ [tɹ:trea, tmə:tme, nt:nt][11];

対応関係の定量評価で有用なのは，音節単位での対応づけであるが，英単語の音節化データベースは存在しない．CMPUDの発音の側の0, 1, 2のON区切りに対応する分割を，語形の方に追加すれば，それが音節の代用単位となると考えた[12]．とすると，(5) にある [ə:a, mɛ:me, ɹɪ:ri, kə:ca, n:n, tɹ:trea, tmə:tme, nt:nt, ...] のような発音と綴りの対を対象語から網羅的に抽出[13]し，それらの出現頻度を数えると，綴りと発音の乖離を定量評価するのに必要なデータが手に入る．また，(5) にあるのは，ON単位の発音と綴りの1-gram対応づけだが，それらを連結して，2-gram対応と3-gram対応が得られる．これらは一般利用可能なデータとしてGitHubで公開している.

3 解析手法と結果

データから得られた，発音の単位 y (e.g, “mɛ”) と綴りの単位 x (e.g., “me”) との対 (x, y) (i.e., “mɛ:me”) に

[3] 本研究で word(s) は lemma(s) ではなく実用語形を指す．例を挙げると，lemmaとしての“run”から，“run”, “runs”, “ran”, “running”, ...のような語形が派生する．実際に知りたいのは，語形の発音だからである．語形の数として4kは，決して大きな数だとは言えない.

[4] これは base-pairs-bundled-r6e.csv として GitHub repository から入手可能．この.csvファイルのfield 1の値は開発データ中のエンコードの回数 (2以上は重複を意味する), field 2の値はCMUPDを使って割り当てた発音記号，fields 3の値は語形を“/”で分割した形 (後述) である.

[5] https://www.sketchengine.eu/

[6] http://www2.cmp.uea.ac.uk/~jrk/conlang.dir/LongmanVocab.html

[7] https://www.rong-chang.com/gsl2000.htm

[8] 前述のように，これはCMUPDのARPABET表記ではなく，それをIPA表記に変換した結果である.

[9] ファイル名は base-ipa-spell-pairs-r6.xlsx

[10] 作業シート W 上に実装されている音変形は幾つかの理由で最適なものとは言えず，かつ誤りを含んでいる可能性がある．このため，後で示す解析結果の妥当性には自ずから限界がある.

[11] tmə/tme のような対応づけ生じるのは，悩ましい問題である．対処法として，treatmentをtreat#mentのように二つの形態素からなると解析し，発音の方も tɹi1t#mə0nt のように分割する事が可能だが，どんな形態素で区切りを入れるかは判断が難しい．それ以前の問題としてhigherをhi/gherと区切るべきかhigh/erと区切るべきか，johnをjoh/nと区切るべきかjo/hnと区切るべきか，など，事前に正解を決められない厄介な問題が他に幾つかあった.

[12] ON分割が音韻論的に自然な単位とは言えないと思うが，2-gramや3-gramを使う事で難点は回避できると考えている.

[13] これには自作のPerl scriptを使った.

ついて，相対確率 α と β を (6) のように定める:

(6) α = 対 (x, y) の出現回数 / 綴り x の出現回数
β = 対 (x, y) の出現回数 / 発音 y の出現回数

これを使って，綴りと発音の対応の評価指標を (7) のように二種類定義できる:

(7) $\gamma = 2\alpha\beta/(\alpha+\beta)$ [実質的に α, β の F 値]
$\delta = \log(\alpha/\beta)$

δ と γ は共に逸脱の程度の評価に使えるが，δ には大きさに加えて方向性があり (δ の値は $\alpha = \beta$ なら 0，$\alpha > \beta$ なら正の値，$\alpha < \beta$ なら負の値)，これが対応の非対称の観点から有用である．

本研究で得られている結果は次の二種類である:

(8) 英語の全データ E の spell-sound 対応の 1-gram, 2-gram の δ 値の分布の比較 (3-gram の結果は割愛)

(9) a. Web サイト 1k most common words[14] で入手したドイツ語の単語に対し，別サイト[15]を使った発音が獲得できた 864 語の spell-sound 対応[16] の 1-gram の δ 値の分布の比較 (2-gram, 3-gram の結果は割愛)[17]

b. データ E 中の Web サイト 1k most common words の英語版に現われている 975 語の spell-sound 対応の 1-gram の δ 値の分布の比較 (2-gram, 3-gram の結果は割愛)

(8) の結果を図 1 と図 2 に示す．(9) の結果を図 3 と図 4 に示す．

図 1–図 4 は指標 $\delta = \log(\alpha/\beta)$ の分布を示している．綴りと発音のズレがないのであれば，$\delta = 0$ である．これらの図はどれも，理想的な一対一対応の状況とはかけ離れており，英単語の綴りと発音の対応は (2) で定義した意味では不規則である[18]．英語の単語の発音と綴りとの乖離は理想の一対一対応とはかけ離れている．

頻度域の同じ語彙を，英語とドイツ語で比較すると英語の状況を相対化できる．図 3 で $\delta = 0$ である対の割合は 0.498 で，図 4 で $\delta = 0$ である対の割合は 0.256 である．英語の綴りと発音の乖離は，それなりに乖離のあるドイツ語に比べても大きい．

[14] https://1000mostcommonwords.com/
[15] https://dad.sprechwiss.uni-halle.de/
[16] 手作業で構築したドイツ語の綴りと発音の対応づけデータ: base-German-ipa-spell-pairs-r1-1k.xlsx
[17] ドイツ語の解析結果: data-German-spell-sound-pairing-r2a-ngram.xlsx
[18] spell-sound 対応の n-gram では，n が増えるにつれて $\delta = 0$ の領域が単調に広がる．これは言語に依存しない傾向である．英語でも 2-gram で領域が広がっている (δ 率が 0.157 から 0.640 に増加).

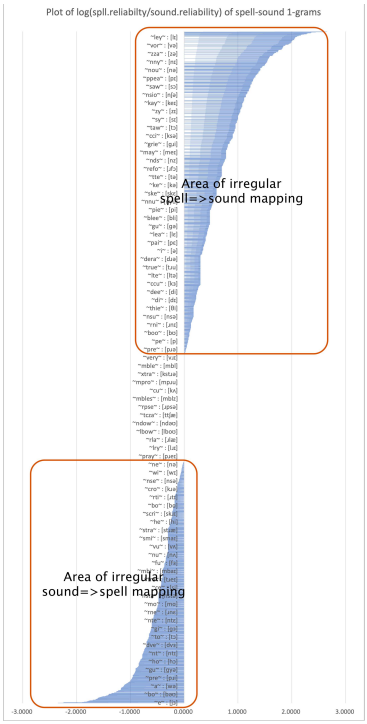

図 1: 英語の **spell-sound** 対 **1-gram** の δ の値のプロット

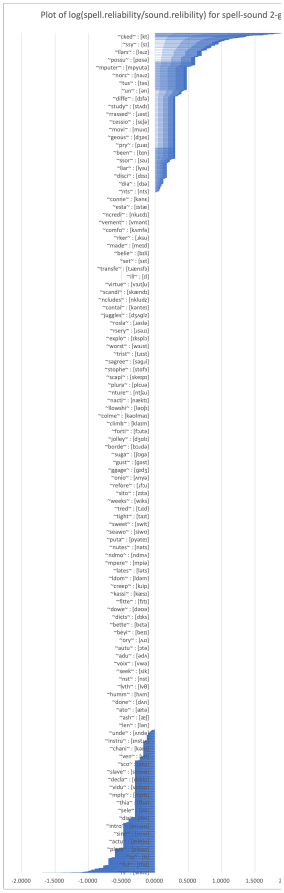

図 2: 英語の **spell-sound** 対 **2-gram** の δ の値のプロット

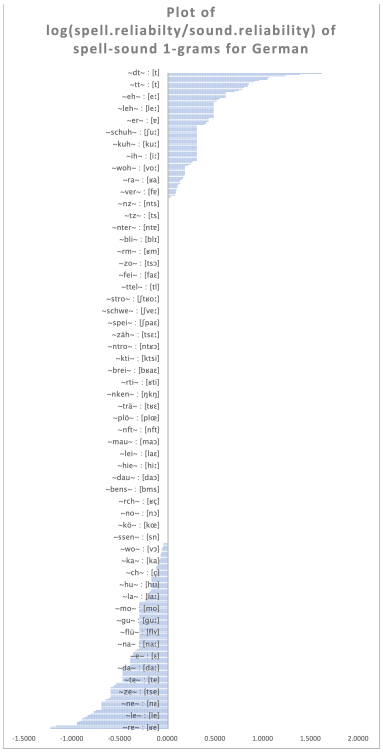

図 3: ドイツ語の **1k most common words** 中の **spell-sound** 対 **1-gram** の δ の値のプロット

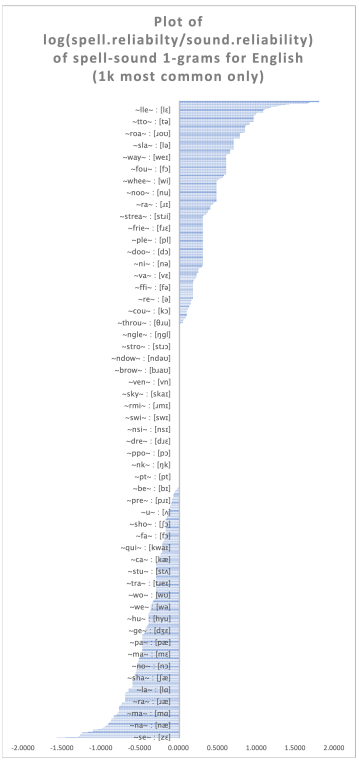

図 4: 英語の **1k most common words** 中の **spell-sound** 対 **1-gram** の δ の値のプロット

4 結論と展望

本研究が明らかにしたのは，二つである．i) 図 1 と図 2 と図 4 が示しているように，英語では，綴りと発音の対応関係が一対一対応から乖離している．ii) 乖離は綴りと発音の対応が透明でない言語[19)]のドイツ語に比べて明らかに大きい．

A) 発音を知っていて綴りを間違うのと B) 発音を知らなくて綴りから間違った発音をするのを区別すると，理論的な予想では $\delta > 0$ なら A 型の誤りを，$\delta < 0$ なら B 型の誤りを誘発するはずである．英語の母語話者は A 型の間違いを犯し，日本人の英語学習者は B 型の間違いを犯しがちだと予想される．この予想を確認したい．

謝辞

認知科学会 40 回大会に応募した際に 2 名の匿名の査読者からコメントを貰った．改稿ではそれから得られるものが多かった．

References

Highly irregular. (n.d.). Retrieved from https://www.readingkingdom.com/pages/english-spelling-is-irregular

Obasi, J. C. (2018). Structural irregularities within the English language: Implications for teaching and learning in second language situations. *J. of English as an International Language, 13*(2.1), 1–14.

Okrent, A. (2021). *Highly irregular: Why "tough," "through," and "dough," don't rhyme and other oddities of the English language.* Oxford University Press.

Patterson, N. (n.d.). *Irregularities.* Retrieved 2023-07-10, from https://www.spellingsociety.org/irregularities-of-english-spelling#/page/1

Wolf, T. (2017). *A descriptive and qualitative-quantitative analysis of the spelling of L1 Spanish-English speaking elementary students* (Unpublished master's thesis). St. Cloud State University.

[19)] 綴りと発音の対応がほぼ透明に近い言語の代表例は，スペイン語，イタリア語，スワヒリ語などである．