外国語多読学習のための獲得語彙の推定分布を考慮した効率的なテキスト選択
Selecting Texts for Extensive Reading Efficiently

江原 遥
Yo EHARA
東京学芸大学　教育学部
Faculty of Education, Tokyo Gakugei University
Email: ehara@u-gakugei.ac.jp

あらまし： 外国語語彙学習では，未習語の辞書引きを避け語義を文意から推測して用例と共に獲得していく多読法がある．多読で読む文章の選択は投機的性格を持つ．簡単すぎれば既習語ばかりで獲得語彙量は少なく，獲得語彙量増加を狙い文章の難度を上げれば，辞書引きなしでは読めず多読の意義を損なうリスクが上がる．本稿ではこの投機的な問題を数理的に定式化し，許容可能なリスクの範囲で予想獲得語彙量最大の文書を選ぶ手法を提案する．**キーワード**：外国語学習，語彙学習支援，多読

1　はじめに

応用言語学分野において，外国語の語彙獲得における付随的学習 (incidental learning) とは，「語彙獲得が主目的ではない活動の中で，偶発的に語彙が獲得されること」[5] である．例えば，テキスト中に出てきた分からない単語の意味を推測しながら外国語の語彙を学習する手法がこれにあたる．付随的学習に対して，いわゆる単語帳を覚える方法など，「語彙獲得を主目的とする活動の中で，語彙を学習すること」を「意図的学習」(intentional learning) と呼ぶ．付随的学習は意図的学習に比べ，外国語の語彙量を増加させるには非効率である一方，文脈の中の語の使われ方など深い理解を促すとされている．本稿では，前述の戦略を機械学習の観点から数理的に定式化し，ある学習者があるテキストを読んだときの付随的学習による獲得語彙量の確率分布を求める手法を提案する．これにより，学習者は読解に失敗しない範囲で獲得語彙量を最大にできるテキストを選択できる．本稿の内容は EDM 2022 poster に採択された [4]．

動機づけのため，例を挙げる．[a,b,c,d] の 4 種の単語からなるテキストを考える．頻度は，それぞれ $[93, 3, 3, 1]$ とする．また，学習者が各単語を知っている確率は [0.9,0.6,0.5,0.2] とする．これは，例えば学習者に対する過去の成績情報などから機械学習等を用い算出する．

この時，応用言語学のテキストカバー率の考え方によれば [5]，テキスト中の 95% 以上の単語を知っていれば，テキストを読むことで知らない単語を学習することができる．具体的に，テキストカバー率が 95% を超える場合を列挙してみよう．この例では，単語 a が多いが，単語 a だけで 95% を超えることはできない．閾値を超えるのは，例えば，{a,b} を知っていて，{c,d} を知らない場合があげられる．知っている単語をだけに注目して書くと，閾値を超える場合は，{a,b}, {a,c}, {a,b,d}, {a,c,d}, {a,b,c,d} の場合である．

さて，この学習者が知っている単語が {a,b} になる場合に注目しよう．{c,d} については知らないと考えていることに注意すると，{a,b} になる確率は，$0.9 \times 0.6 \times (1-0.5) \times (1-0.2)$ と計算できる．この時，前述の応用言語学の知見 [5] は，この学習者がこのテキストを読むことで，**知らない単語 $\{c, d\}$ を文意を通じて獲得できる**と考えられるので，この確率はこの学習者が単語 {c,d} を新たに獲得できる確率ともみなせる．こうして，すべての場合について考え，この学習者がこのテキストを読むことで [a,b,c,d] の各単語を獲得できる確率を集計していくと，[0,0.18,0.27,0.576] になる．このうち，単語 a については獲得できる確率が 0 になっているが，これは，単語 a を知らなければそもそもテキストが読めず獲得が起こらないことを示す．

2　提案する定式化と解法

前節の内容を定式化しよう．

今，I 種類の語彙 $\{v_1, \ldots, v_I\}$ を考え，注目しているテキスト中の v_i の個数を n_i とする．また，注目している学習者が v_i を知っている確率を p_i で表す．この時，テキストカバー率の閾値を τ とする（前節の例では $\tau = 0.95$）．この時，テキストカバー率が閾値を超える確率は，次のように表せる．まず，テキスト中の総単語数は $N = \sum_{i=1}^{I} n_i$ と表せる．また，学習者が単語 v_i を知っている場合に 1，知らない場合に 0 になる次の確率変数を考える（ただし，$\{Z_1, \ldots, Z_I\}$ は互いに独立）．

$$Z_i \sim \text{Bernoulli}(p_i) \tag{1}$$

この時，学習者が知っている単語のテキスト中の出現数は $\sum_{i=1}^{I} Z_i n_i$ であるから，テキストカバー率は $\frac{\sum_{i=1}^{I} Z_i n_i}{N}$ と表せる．したがって，テキストカバー率が閾値を超える確率は $P(\sum_{i=1}^{I} Z_i n_i \geq N\tau)$ と表せる．

学習者がテキストを読む付随的学習により語 v_i を新たに獲得する確率は，次のように定式化できる．テキストを読む付随的学習が起こるためには，テキストが読める必要があるので，テキストカバー率が閾値を超えていなければならない．さらに，語 v_i が新たに獲得されるためには，学習者は語 v_i を知らないことが必要である．したがって，この確率は，$P(Z_i = 0, \sum_{i=1}^{I} Z_i n_i \geq N\tau)$ と表せる．この確率を q_i とおく．さらに，このテキストからの付随的学習による獲得語数の分布を求めたい．$A_i \sim \text{Bernoulli}(q_i)$ と置くと，獲得語数の確率変数 A は，$A = \sum_{i=1}^{I} A_i$ と表せる．したがって，A の確率分布を求めれば，獲得語数の分布も求まる．ただし，$\{A_1, \ldots, A_I\}$ は互いに独立とする．

テキストカバー率が閾値を超える確率 $P(\sum_{i=1}^{I} Z_i n_i \geq N\tau)$ や，獲得語数の分布 $A = \sum_{i=1}^{I} A_i$ を求めるためには，**異なる成功確率**を持った互いに独立な二項分布の和からなる確率変数の分布を求める必要がある．成功確率が等しければ二項分布の和は再生性を持つが，成功確率が異なるため，この和は二項分布にはならない．こうした分布は，**ポアソン二項分布**と呼ばれる．$P(\sum_{i=1}^{I} Z_i n_i \geq N\tau)$ については，動的計画法を解くことで求める方法を筆者が過去に提案している [3]（第 44 回教育システム情報学会大会奨励賞受賞）．簡潔に言えば，$\sum_{i=1}^{I} Z_i n_i$ が整数であることを利用して，$N\tau$ 以上という条件を，「$\{n_1, \ldots, n_I\}$ の部分和が丁度いくつ」という部分和問題に帰着させる．この部分和問題を，「$\{n_1, \ldots, n_i\}$ までの数で丁度いくつになるものを作れる確率」からなる DP テーブルによる動的計画法で解ける．今回は，それに加えて，$P(Z_i = 0, \sum_{i=1}^{I} Z_i n_i \geq N\tau)$ を求める必要がある．こちらは，動的計画法の DP テーブルの各セルに，そのセル時点での $\{Z_1, \ldots, Z_I\}$ の確率値の集計値を記録する拡張を施すことで計算した．なお，ポアソン二項分布は，平均と分散を求めるだけであれば，$\sum_{i=1}^{I} p_i$ が平均，

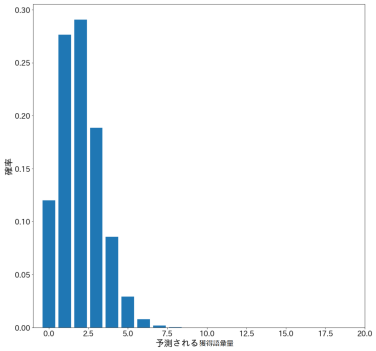

図 1: 予測される獲得語彙量の分布の例.

$\sum_{i=1}^{I} p_i(1-p_i)$ が分散となる.

3　実験

語彙テスト結果については，著者がクラウドソーシングを用いて過去に公開したデータがある [2]．具体的には，クラウドソーシング上の学習者（TOEIC の受験経験があるものに限定）100 人に，100 問からなる語彙サイズ計測用の単語テスト Vocabulary Size Test (VST)[1] を受けてもらった結果のデータセットである．VST は多肢選択型のテストであり，**英文中に埋め込まれた**単語の言い換えとして適切な選択肢を 4 つの選択肢の中から選択するテストである．

単語テストの結果を使って，各学習者が所与の単語を知っているかどうかを判別する確率的識別器を作成し，この確率値を式 1 における p_i として用いた．単純なロジスティック回帰を用いて識別器を構成した．特徴量としては，COCA コーパスの頻度，British National Corpus (BNC) コーパスの頻度を用いた．ただし，頻度は – log(頻度) の形に直して特徴量として用いた．テキストとしては，Brown Corpus を用いた．テキスト長さが実験結果に影響しないように，Brown Corpus の 500 件の各テキストのうち，先頭から 300 語を切り出し，実験に用いた．この 500 件のテキストから，ある学習者の付随的学習に適したテキストを選択することが，我々の目的である．付随的学習はテキストを読める必要があるので，好成績の学習者に起こりやすい．まずは，最も成績の良かった学習者（VST で 96 問正解）を対象に実験を行った．図 1 にこの学習者があるテキストを読んだ時の獲得語彙量の分布を 1 つ示す．図 1 より，獲得語彙量の分布には幅があり，単純に期待値が高いテキストを選べばよいわけではないことがわかる．

各テキストを読む際に予測される獲得語彙量の期待値と分散を同時に考慮するため図 2 に図示した．各点は Brown Corpus の各テキストである．獲得語彙量を学習者にとっての利得と考えると，獲得語彙量の期待値が同じであれば，できるだけ獲得語彙量の分散が少ないテキストを選択する方が，確実に語彙を増やせるので，学習者にとっては得となる．すなわち，図 2 の左上部分が，この学習者にとって最も効率的に付随的学習を行える文書群である．このように，図 2 の縦軸は利得，横軸の獲得語彙量の分散は**リスク**とみなせ，図 2 は，経済分野で多用されるリスクとリターンの関係図とみなせる．

図 2 のようなリスクとリターンの関係図においては，左上部分が最も低リスクで利得を増やせる選択であり，この部分を**効率的フロンティア**という．図 2 では，500 件あった選択肢の中から，凸包を用いて効率的フロンティアに属するテキスト 5 件が選択された．すなわち，学習者の付随的学習に適したテキストを 1/100 に絞ることができた．この 5 件の中でどのテキストを選択するかは，学習者がどの程度のリスクをとって語彙量を増やしたいかによって変わるので学習者に任せる方法が一案である．

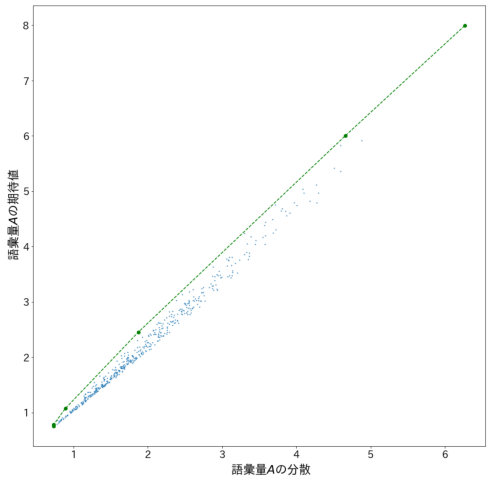

図 2: 各テキストの獲得語彙量分布の期待値と分散.

学習者によって効率的フロンティア曲線は変わるが，多くの学習者に薦められるテキストはあるのだろうか？100 人のうち，成績の良い 30 人に対して，同様に各学習者が各テキストを読んだ場合に予測される獲得語彙量の分布から効率的フロンティアを求めた．その結果，10 人以上で，効率的フロンティアに選ばれたテキストが 7 件あり，最も多いもので 14 人の効率的フロンティアに含まれたテキストがあった．このテキストは，具体的には “In the imagination of the nineteenth century the Greek tragedians and Shakespeare stand side by side , their affinity transcending all the immense contrarieties of historical circumstance , religious belief” で始まるテキストであり，“contrarieties” という難しい語が一部使われている他は，“belief” など中級学習者なら習得している語で構成されていることが見て取れる．この結果から，効率的フロンティアに含まれるテキストは比較的安定しているように思われる．

4　おわりに

本研究では，付随的学習による語彙獲得に適したテキストを選択するため，応用言語学の知見に基づき，個々の学習者・個々のテキストに対して付随的学習によって獲得される語彙量の推定値を算出する手法を提案した．獲得語彙量を学習者にとっての利得と考えることで，金融工学などで使われる効率的フロンティアの考え方を語彙学習支援の研究に導入した．個々の学習者に適応的に効率的なテキストを求められることはもちろん，多くの学習者の効率的フロンティアに含まれることで一般に付随的学習に適したテキストの存在を実験的に示した．

謝辞　本研究は JST 戦略的創造研究推進事業 ACT-X（JPMJAX2006）の支援を受けた．

参考文献

(1) David Beglar and Paul Nation. A vocabulary size test. *The Language Teacher*, Vol. 31, No. 7, pp. 9–13, 2007.
(2) Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, 2018.
(3) Yo Ehara. テキストカバー率の確率的拡張に基づく語彙テストのみからの個人化読解判定. 2019.
(4) Yo Ehara. Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary. In *Proc. of EDM (poster)*, 2022.
(5) 中田達也. 英単語学習の科学. 研究社, 2019.