順序ロジットモデルに基づいた 英単語の習得困難度の推定とその要因の分析

田中省作[*1]　木村 恵[*2]　依田みづき[*3]　八島 等[*4]

*1 立命館大学文学部　*2 獨協大学外国語学部　*4 東京都立城東高等学校
*3 Victoria Univ. of Wellington, School of Linguistics and Applied Language Studies

1　はじめに

現在，我々は主に中高生を対象とした，英語学習（習得）の診断テスト EDiT(English Diagnostic Test) の開発を行っている [3]．EDiT は文法・語彙・音声の 3 つの診断テストから成る．その中で語彙の診断テストは，英単語の習得が十分期待される段階（**習得困難度**）と学習者の英語学習段階に基づいて診断を下す．この英単語の習得困難度は，特に中高であれば教科書での語彙の使用状況によって容易に規定できるように思われる．しかし実際には，このような情報と，中高学習者の語彙習得状況それに近い英語教師らの経験・実感とは完全には一致しない．そこで本研究では，まず英単語の習得困難度に関係する要因（**習得困難要因**）を，(1) 単語そのもの，(2) 教科書，(3) 学習が保証される基本語彙の 3 つの観点から，網羅的に設定する．そして，それらを数量化した情報と，語彙の習得状況に関するデータから，英単語の習得困難度の推定モデルを構築する．推定モデルには順序ロジットモデルを採用し，AIC に基づいて習得困難要因の絞り込みを行う．実験の結果，小規模データではあるが，ベースラインを超える精度が得られた．

2　英単語の習得困難度と要因

2.1　英単語の習得

EDiT の語彙の診断テストは，中高での活用を念頭に 1 セット 15 分，文法・語彙・音声の計 45 分程度（授業 1 コマ）で構成される．時間的制約から，この診断テストにおける英単語の習得は，英語学習おける最低限の保証を与えるという立場で『その単語の意味を知っている』と考えている[1]．

しかし，このように英単語の習得を単純化したとしても，学習者が英単語を習得する段階（習得困難度）を正確に規定することは難しい．現状では，高校生を対象とした僅かなパイロット・テストや，英語教師らの直感に頼らざるを得ない．この診断テストの作成には，習得困難度が付与された大規模な英単語リストを前提としている．そこで本研究では，まず英単語の習得困難度に関係する要因（習得困難要因）を設定し，習得困難度の自動推定，あるいは人手による習得困難度の付与支援のための方法論を検討する．

2.2　習得困難要因

英単語の習得困難要因を，第二言語習得研究における知見，アンケートやパイロット・テスト，英語教師らの議論に基づき選定した．この段階では，要因間の相関は考慮せずに列挙している．その概要を表 1 に示す．特徴的な点は，単語そのものに関わる情報以外に，(1) 中高の英語学習を強くコントロールする教科書に関わる情報，(2) 当該単語と基本語彙に関わる情報を考慮した点である．

(2) の基本語彙とは，中高の教科書で必ず学ぶ語彙である．教科書といっても出版社や編纂者が異なれば，含まれる語彙は大きく変わる．高校の教科書にいたっては，中学の教科書にはない 3 つのレベルが設定されている[2]．ここでは，平成 14 年度版検定

[1] [5] によれば，単語の知識は Form (spoken, written, word parts), Meaning (form and meaning, concept and referents, associations), Use (grammatical functions, collocations, constraints on use) と細分化され，英単語の習得は多角的に考える必要がある．この考え方に基づけば，語彙の診断テストでも単語のコロケーションや意味制約などの単語知識の深さや，/æpl/を “apple” と認識する音声的な側面も勘案しなければならない．

[2] 判型で区別される．

表 1: 習得困難要因

単語に関わる情報
品詞
多義性
抽象性
親密性
大きさ
教科書に関わる情報
教科書に出る時期
高校教科書のレベル
教科書での頻度
教科書での散らばり*1
基本語彙に関わる情報*2
形態の類似性
音声の類似性
意味の類似性
対訳の類似性

*1 教科書のどのレッスン・ユニットでも出る単語なのか，特定のレッスン・ユニットだけで出るようなものか.

*2 基本語彙に形態・音声・意味・対訳の面で当該単語と類似している単語があるか（または，どの程度類似したものがあるか）.

済の中高教科書から，中学教科書 3 冊の本文パートで，どの教科書が採択されていても必ず学ぶ[3]402 語と，中位レベルの高校教科書 4 冊で同様の 431 語，計 833 語を基本語彙としている[4]．なお，単語は [4] にならい，品詞で細分化し，規則活用するものについては原形（asked_V → ask_V, taller_AJD → tall_ADJ），不規則活用するものについてはそのまま（did_V，better_ADJ）取り扱う.

3 順序ロジットモデルに基づいた順序付き名義カテゴリの推定

4 節で示す実験では，英単語の習得困難度を 5 段階に設定している．レベル 1 が中 1，レベル 2 が中 2,3，レベル 3 が高 1，レベル 4 が高 2，レベル 5 が高 3 程度の学習者が十分に習得できる単語，という具合である．この『レベル y』は順序関係がある名義変数である．順序関係がある名義応答を予測するモデルの一つに，計量経済学や金融工学で会社の格付や信用リスク評価などに利用される順序ロジットモデルがある [1, 2].

カテゴリ $y_1, y_2, \cdots, y_k$ には順序関係があり，応答変数 Y はカテゴリのインデックスを表すとする．$x_1, x_2, \cdots, x_\ell$ を説明変数とすると，順序ロジットモデルにおいて，$P(Y \leq j)$ のロジットは次のように与えられる.

$$\mathrm{logit}[P(Y \leq j)] = \alpha_j + \sum_{i=1}^{\ell} \beta_i x_i$$

$\alpha_j (1 \leq j \leq k-1), \beta_i (1 \leq i \leq \ell)$ は定数で，β_i はカテゴリ j には依存しない．なお，ロジット p は，

$$\mathrm{logit}[p] = \log \frac{p}{1-p}$$

である.

ここで，$x_1, x_2, \cdots, x_n$ のデータに対するカテゴリの推定問題を考えると，

$$P(Y \leq j) = \frac{\exp\{\alpha_j + \sum_{i=1}^{\ell} \beta_i x_i\}}{1 + \exp\{\alpha_j + \sum_{i=1}^{\ell} \beta_i x_i\}}$$

となり，

$$\begin{aligned} \widehat{Y} &= \operatorname*{argmax}_j P(Y = j) \\ &= \operatorname*{argmax}_j \{P(Y \leq j) - P(Y \leq j-1)\} \end{aligned}$$

と決定される.

サイズが N の標本に対する尤度は，次式のように与えられ，α_j, β_i は最尤推定される.

$$L = \prod_{j=1}^{k} \prod_{i=1}^{N} \left\{ P(Y^{(i)} = j \mid x_1^{(i)}, x_2^{(i)}, \cdots, x_\ell^{(i)}) \right\}^{I_j^{(i)}}$$

ただし，$Y^{(i)}$ は標本で i 番目のデータのカテゴリ，$x_j^{(i)}$ は i 番目のデータの x_j の値，$I_j^{(i)}$ は二値で i 番目のデータのカテゴリが j のとき 1，そうでない場合は 0 である.

4 習得困難度の推定実験

4.1 データ

中高教科書の英単語からランダムに 266 語を抽出し，英語教師 7 人が習得困難度を 3.1 節冒頭で述べ

[3] 3 年間一貫して同じシリーズの教科書を使うことを前提とする.

[4] 中学は採択率が特に高い NEW CROWN, NEW HORIZON, SUNSHINE, 高校は中位レベルで I/II/Reader が揃っている DREAM MAKER, ENGLISH21, PHOENIX, SPECTRUM を用いた.

た 5 段階で個別に付与した．5 人以上で習得困難度が一致した英単語 162 語を実験データとした．

4.2 習得困難要因の数量化

2.2 節・表 1 で挙げた要因を，各種辞書（EDR 電子化辞書（英語コーパス・日本語単語辞書・英語単語辞書・英日対訳辞書），Roget シソーラス）と表 2 の要領で数量化した．

4.3 習得困難要因の選択（モデル選択）

最適な順序ロジットモデルの選択，すなわち習得困難要因（説明変数）の選択を AIC に基づいて行った．習得困難要因の全ての組合せ 65,535 通りのモデルを構築し，AIC が最小のモデルを採用する．k 個の応答カテゴリ，ℓ 個の説明変数で構成される順序ロジットモデルの AIC は，次式で計算される．

$$\mathrm{AIC} = -2\log L + 2(\ell + k - 1)$$

ただし，L は最大尤度である．

4.1 節のデータを用いた実験の結果，AIC が最小となったモデルは,『抽象性』『親密性』『中学教科書における（平均相対）頻度』『音声の類似性』『意味の類似性』の 5 つを説明変数とするものであった．

4.4 推定の精度

4.3 節で採用したモデルの推定精度について述べる．精度は次のような手順で算出した．データ 162 語のうち 161 語を学習データとして α_j, β_i を推定し，その後，残された 1 語の習得困難度を推定する．これを全ての単語がテスト・データとなるように，162 回繰り返し行った．

推定の結果を表 3 に挙げる．正しく推定された単語が 113 語（表内ボールドの数字）で，正解率は 69.8%である．最も単語が多い習得困難度はレベル 1 なので，ベースラインは 53.1%と考えることができる．データ規模が小さく有意とは言えないが，ベースラインを上回る結果を得た．また，$\widehat{P}(Y=j)$ の上位 2 位までのカテゴリに正解が含まれる割合は 92.6%，上位 3 位では 99.4%となった．

表 3: 推定結果

$Y\backslash\widehat{Y}$	1	2	3	4	5
1	**73**	6	1	0	0
2	13	**6**	2	0	0
3	0	1	**30**	2	1
4	0	0	12	**2**	3
5	0	0	4	4	**2**

4.5 習得困難要因の分析

モデルの β_i から，要因が習得困難度に与える影響を読み取ることができる．採用したモデルの各習得困難要因に対する β_i の正負は，表 4 の通りである．表内の順序は説明変数を標準化して，β_i を求めたときの絶対値の降順である．

$\beta_i > 0$ の場合，x_i が大きくなるにつれ小さなカテゴリの確率が高くなり，逆に $\beta_i < 0$ の場合，x_i が大きくなるにつれ大きなカテゴリの確率が高くなる，という関係がある．したがって，次のような傾向が読み取れる．ただし，今回の実験のデータ規模は小さく，現時点ではあくまでも試みとして解釈するものである．

- 中学教科書に頻出する単語は習得が易しい
- カタカナ対訳が定着している単語は習得が易しい
- 抽象的な単語は習得が難しい
- 当該単語と発音が類似した基本単語が**ない**場合，習得は難しい
- 当該単語と類義語関係にある基本単語が**ある**場合，習得は難しい

以上は,『発音の類似性』を除き，[3] で従来研究の知見から予想された傾向と一致している．このような，要因が習得困難度に与える影響については，専門家の間でも議論が分かれることも多い．このようなデータ主導の分析が，そういった議論の検証の一助となることが期待される．

表 2: 習得困難要因の数量化

要因の候補	数量化の概要
品詞	ダミー変数で表現.
多義性	語義頻度分布のエントロピー（語義頻度分布が無い場合があるので，当該単語の語義頻度分布の有無を表すダミー変数と組で扱う）.
抽象性	UCREL Semantic Analysing System で “general and abstract terms” とされる場合 1，そうでなければ 0.
親密性	カタカナ対訳が存在し，なおかつそのカタカナ対訳が日本語辞書の項目となっている場合 1，そうでなければ 0.
大きさ	(1) 文字数，(2) シラブル数.
教科書に出る時期	全ての教科書で当該単語が出揃う学年（中 1 → 1，$\cdots$，中 3 → 3，高 1 → 4，$\cdots$，高 3 → 6，それ以外→ 7).
高校教科書のレベル	教科書のレベルごとに当該単語の頻度を求め，χ^2 検定により特徴的に頻出しているレベルをダミー変数で表現（$\alpha = 0.05$）.
教科書での頻度	(1) 中学教科書一冊あたりの平均相対頻度，(2) 高校教科書一冊あたりの平均相対頻度.
教科書での散らばり	(1) 当該単語を含む中学教科書一冊あたりの平均相対パート数，(2) 当該単語を含む高校教科書一冊あたりの平均相対パート数.
形態の類似性	(i)3/4 以上の部分文字列が一致し，(ii) 単語の文字数が 0.8-1.2 倍，(iii) 前または後 3 文字が一致，の 3 条件を満たす基本単語と当該単語との編集距離の最小値（条件を満たす基本単語が無い場合があるので，比較対象の有無を表すダミー変数と組で扱う）.
音声の類似性	当該単語の発音記号列と基本単語の発音記号列の編集距離の最小値.
意味の類似性	英語シソーラスで当該単語と類義語関係となる基本単語が存在する場合 1，そうでなければ 0.
対訳の類似性	当該単語の英日対訳を共有する基本単語が存在する場合 1，そうでなければ 0.

表 4: 習得困難要因と β_i の正負

$\beta_i > 0$	$\beta_i < 0$
音声の類似性	中学教科書での頻度
意味の類似性	親密性
抽象性	

5 おわりに

順序ロジットモデルと AIC を用いて，複数の習得困難要因を考慮した英単語の習得困難度推定モデルを構築した．要因を数量化する過程，データの作成・収集方法そして量と，検討すべき点は多々残されている．また，β_i の正負だけではなく，より詳細に，個別データの振る舞い，習得困難要因の複合的な影響を観察することも今後の課題である．データに関しては，2 月から実際に中高生に対して英単語の習得状況の調査が計画されており，そこで得られた実データを用いた実験および分析を進める予定である．

参考文献

[1] Agresti, A.: Categorical Data Analysis, Hoboken, N. J.: Wiley (2002).

[2] Fienberg, S. E.: The Analysis of Cross-Classified Categorical Data, The MIT Press (1980).

[3] 金谷 憲編: 英語診断テスト開発への道, 英語運用能力評価協会（ELPA）(2006).

[4] Kimura, M., Tanaka, S. and Tomiura, Y.: Tracing Japanese EFL Leaners' Development in Productive Vocabulary, *Proc. of the NICT JLE Corpus Symposium*, pp. 54-71 (2005).

[5] Nation, I. S. P.: Learning Vocabulary in Antoher Language, Cambridge University Press (2001).

[6] 投野由紀夫編: 英語語彙習得論, 河源社 (1997).