様式 C-19

# 科学研究費補助金研究成果報告書

平成 21 年 6 月 29 日現在

研究種目：若手研究(B)
研究期間：2007～2008
課題番号：１９７２０１４９
研究課題名（和文）
順序尺度としての学年に基づいた日本人の英語リーダビリティ推定式の開発
研究課題名（英文）
Construction of the formulae estimating the readability of English texts for Japanese learners of English on the basis of grades on ordinal scales
研究代表者
田中　省作（TANAKA, SHOSAKU）
立命館大学・文学部・准教授
研究者番号：００３２５５４９

研究成果の概要：
本研究は，英文書のリーダビリティの指標としての「学年」を順序尺度として捉え，日本人英語学習者のためのリーダビリティ推定式を構築した．推定式の数理モデルに順序ロジットモデルを採用し，日本人英語学習者の英文書に対するリーダビリティ意識を調査した上で，日本の中高検定英語教科書に基づきリーダビリティ推定式を具体的に構築，実験でその有効性を確認した．構築した推定式は Web 上で公開している．

交付額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2007 年度 | 2,000,000 | 0 | 2,000,000 |
| 2008 年度 | 1,400,000 | 420,000 | 1,820,000 |
| 年度 | | | |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,400,000 | 420,000 | 3,820,000 |

研究分野：知能情報学、言語学
科研費の分科・細目：言語学・外国語教育
キーワード：リーダビリティ，言語教育学，統計数理，英語

## １．研究開始当初の背景

英語教育分野を中心にリーダビリティの数量的な推定に関する研究が進んでおり，その主流は文書の種々の情報（語彙や文長など）を線形和のような数式によって総合し，リーダビリティを算出するものである．このような研究の成果としては，Dale-Challの公式やFleschの公式などが有名で，実際にパソコンなどの文書作成アプリケーションで利用することもできる．教育現場では，教材選定やテスト問題の検証などの基本的な道具の一つとしても役立っている．しかし，このような数式に基づくリーダビリティ研究は，主に次のような2つの問題を抱えていた．

(1) リーダビリティの指標である「学年」を量的に捉えていたこと
先行研究のほとんどが，リーダビリティ推定を多変量解析の一つである重回帰モデルに帰着し，推定式を構築している．多くの研究がリーダビリティの指標として「学年」を採用しているが，その目的変量としての「学年」を素直に量として取り扱い，重回帰モデルを

適用する．これを認めたとすると，「学年に対してリーダビリティも一定等間隔に上昇する」という不自然な仮定を置くこととなる．

(2) 母語話者のリーダビリティの順序関係が，非母語話者である日本人英語学習者において必ずしも成り立つとは限らないこと

先行研究でも指摘されている，日本人英語学習者に対するリーダビリティ研究の重要な問題である．日本人の英語学習過程は母語話者とは大きく異なり，特に教科書などの身近な教材などの影響が強いことが予想されており，リーダビリティに関わる要因も異なる可能性がある．

本研究は，以上のような問題の解決を図る．

## ２．研究の目的

本研究の目的は前項で述べた 2 つの問題を念頭に，リーダビリティ指標である「学年」を順序尺度として捉えた日本人英語学習者のための英文書リーダビリティ推定式の構築を行うことである．そのために，次のような小テーマを設定した．

(1) リーダビリティ推定式の数理モデルの再考

(2) 日本人英語学習者のリーダビリティに関する基本データの蓄積と分析

(3) 日本人英語学習者のための英文書リーダビリティ推定式の構築と評価

(4) リーダビリティ推定式の Web での公開

また，このリーダビリティ推定式から得られる学年とリーダビリティ間の（曲線的）関係から，教育的示唆などの知見を得ることも副次的な目的である．

## ３．研究の方法

本研究のリーダビリティ推定式の構築には，単に推定式の数理的な枠組みを与えるだけではなく，具体的な推定式の構築と検証を相互に繰り返すためのデータが必須となる．研究者間で共通して使用できるようなリーダビリティ研究のためのデータは，研究代表者が知り得る限りでは存在しない上に，そもそも本研究では対象を日本人英語学習者（以後，日本人と記す）に特化しているため，より特殊なデータが求められる．したがって，このようなデータを地道に蓄積する必要があった．そこで本研究では，(1)日本人のリーダビリティに関わるデータの蓄積と分析，(2)リーダビリティ推定式（以後適宜，推定式と記す）の数理的検討と具体的な推定式の構築・評価，という 2 つの中テーマを相互に推進した．

[2007 年度]

(1) 日本人のリーダビリティに関する基本データの蓄積と分析

①日本の中高検定英語教科書の分析

本研究ではリーダビリティ推定式の指標を日本の学校教育制度における「学年」とする．そこで日本の中高検定英語教科書（以後，中高教科書と記す）における文章を基準資料とした．日本人のリーダビリティ意識に関わる要因を吟味するためにも，これらの基本的な事項（語彙・文長といった粗い構造など）の分析を行った．

②日本人向けの親密語リストの作成

Dale-Chall の公式では，リーダビリティを文書の平均文長（単語数）と難語率（Dale リストとよばれる親密語リストの語彙が文章に占める割合）に基づいて推定している．この両要因は，従来のリーダビリティ研究における数多くの要因の提案と取捨選択のなかで，現在も精度と予測性能の観点で優れたものであることが報告されている．本研究でも最も基本的な要因については Dale-Chall の公式を踏襲した．そして，Dale リストに相当する日本人向けの親密語リストを主要な中高教科書に基づき作成した．教科書シリーズによって語彙の導入時期・使用状況が異なるので，どの教科書シリーズを使っても必ず学ぶ（出揃う）時期に着目し，7 種類の親密語リストにまとめた．

③英単語の習得時期の推定

教科書に含まれる語彙は非常に限られており，前項の親密語リストで捉えられる英文書の語彙に関する情報は断片的なものである．そこで，日本人にとっての英単語の習得時期（≈習得が期待される時期）を，その単語の各種情報（品詞や多義性，教科書での出現時期や頻度，基本単語との形態や音声の類似性など）から自動的に決定する枠組みを開発した．

(2) リーダビリティ推定式の数理モデルの再考

「学年」を順序尺度として取り扱うために，従来の重回帰分析に替わる数理モデルの導入を検討した．最初に挙がったものが順序ロジットモデルである．順序ロジットモデルは，順序関係が仮定されたカテゴリに対して，特

定のカテゴリ以上の累積確率分布を推定するもので，数理的な健全性を保ちつつベイズ決定法などを組み合わせることで文書に対するリーダビリティ指標としての「学年」を決定することができる．

(3) 学校文法の項目の自動検出のための予備調査

代表的なリーダビリティ推定式では文書中の学校文法に関する情報は，その要因として考慮されていない（過去，採用を検討した試みはいくつかある）．日本人の英語学習過程を勘案すると，こういった学校文法の諸要素も日本人のリーダビリティ意識に強い影響を与えていることが予想される．そこで，本研究ではそれらを推定式に積極的に取り込むための予備調査を行った．

[2008 年度]

(1) 英語圏のリーディング教材の蓄積と日本人のリーダビリティ意識の調査

母語話者と日本人のリーダビリティ意識が，どういった性質の英文書同士では同順/交差するのかを調査するために，まず母語話者に対する学年レベルが付与された英語圏のリーディング教材の蓄積を行った．そして，それらのデータに対して日本人のリーダビリティ意識の調査を行った．具体的には，物語文・説明文のジャンル別に日本人大学生を被験者としたリーディング実験を行った．実験では連続した 3 学年各 1 文書の計 3 文書を 1 組として読解し，その後 3 文書における絶対的な難易度・相対的な順位を付与した．

(2) リーダビリティ推定式の構築と評価

前項までに蓄積したデータを活用して，具体的なリーダビリティ推定式を構築し，性能評価・検証を行った．

①リーダビリティ推定式の評価

・日本の英語教科書や副読本での評価

本研究で構築したリーダビリティ推定式と日本人向けに再学習した Dale-Chall の公式のそれぞれで，日本の教科書や副読本中の英文書に対してリーダビリティを推定し，その精度を比較・検討した．

・文書のジャンル（物語文・説明文）での評価

2008 年度(1)のデータに対して，ジャンルごとに本研究および既存のリーダビリティ推定式（母語話者向け 7 種・日本人向け 2 種）が付与する順位関係と，母語話者・日本人のリーダビリティ意識の順位関係を比較・検討した．

②リーダビリティ推定式の数理的枠組みの拡張

2007 年度導入した順序ロジットモデルは比例オッズ性という強い制約を課す．この比例オッズ性を仮定しない数理モデルとして拡張逐次ロジットモデルの導入を検討した．

③リーダビリティ推定式の公開

本研究で構築したリーダビリティ推定式を，インターネット経由で誰でも自由に使えるように Web で公開した．

## ４．研究成果

本研究では，主に次のような研究成果を挙げた．

(1) 日本人のリーダビリティ意識に関わる各種基本データの蓄積と分析

①日本の中高検定英語教科書の分析

日本の中高教科書を，リーダビリティ推定式構築のための基礎データとして蓄積すると同時に，特に語彙とリーダビリティの観点から分析し，その詳細と教育的示唆を示した．

②日本人向けの親密語リストの作成

日本人のためのリーダビリティ推定式を構築するにあたり，単語に関する新しい情報として，日本の中高教科書において全てのシリーズで出揃う学年を算出し，学年別の親密語リスト（中 1: 110 語，中 2: 137 語，中 3 : 140 語，高 1: 182 語，高 2: 235 語，高 R: 472 語）を整備した．

③英単語の日本人での習得時期の推定

英単語のさまざまな情報を数量化し，各英単語に対する日本人の習得時期を統計的に推定する方法を提案した．英語教師らによって作成された習得時期付き英単語 162 語に対して実験を行い，ベースラインを上回る 69.8%の推定精度を確認した．この精度は習得時期を完全自動で推定するには不十分であるものの,上位 2 位までの精度は 92.6%で，英単語データの習得時期の付与支援などには有効であろう．今後のデータの蓄積や単語情報の数量化の改良によって，さらなる精度向上も期待できる．

また，この実験結果から，単語情報のうち「基本単語との音声の類似性」「中学教科書での頻度」「カタカナ対訳の有無」が当該英単語の習得時期に強く関わることが示唆され，おおむね従来の語彙研究における見解と一致した．

④母語話者と日本人のリーダビリティ意識の調査

母語話者を対象とした英語圏の学年レベ

ル付きリーディング教材を，物語文・説明文のジャンル別で日本人大学生延べ 45 名を被験者としたリーディング実験を行い，母語話者・日本人の間のリーダビリティ意識に関するデータを作成した．このような母語話者・日本人の両者のリーダビリティ意識に関わるデータは稀有である．このデータを分析した結果，説明文においては母語話者と日本人の意識はほぼ同順，一部の物語文については全く保存されない，という傾向がうかがわれた．さらに，このような意識の差を精査すると，特に物語文の場合については，「テキストの量（≈文数)」が（非線形ではあるが）大きな影響を与えていることが明らかとなった．リーダビリティ推定式で，このようなテキスト量を考慮しているものは少なく，今後の重要な観点であると考えている．

(2) 順序尺度としての学年を指標としたリーダビリティ推定式の構築

①リーダビリティ推定式の数理モデルの再考

従来の重回帰分析モデルに替えて，順序ロジットモデルを導入し，リーダビリティ指標である「学年」を順序尺度として解釈し，リーダビリティ推定式を構成する枠組みを示した．また，要因の選択については，実際に学習データからリーダビリティ推定式を構築する際に，AIC などのモデル選択に帰着することとした．現データ量に相応の最適な要因の組み合わせを決定することができ，今後のデータの拡充に合わせて精密化していくことも可能である．

さらに，順序ロジットモデルにおける比例オッズ性を仮定しない数理モデルとして，拡張逐次ロジットモデルを検討した．拡張逐次ロジットモデルについては，現在のデータ量に対してモデルの自由度が高すぎるため，具体的な推定式の構築は見送った．今後のデータの蓄積に合わせて部分的にでも現モデルから拡張逐次ロジットモデルへ複雑化していけば，より高精度な推定式の構築が期待される．

②リーダビリティ推定式の構築・評価

日本の中高教科書データを学習データとして，順序ロジットモデルと AIC に基づいたリーダビリティ推定式を実際に構築した．同データで日本人向けに再構築した Dale-Chall の公式と推定精度を比較し，本研究の推定式が 51.9%，再構築した Dale-Chall の公式+Nearest-neighbor 法が 43.9%という精度で，わずかではあるものの本研究の推定式の方が有意に優れていることを確認した．

また，前項 (1)-④の日本人のリーダビリティ意識を付与した英語圏のリーディング教材に対して，既存の推定式（母語話者向け 7 種・日本人向け 1 種）と本研究の推定式での順序関係を算出し，日本人のリーダビリティ意識との関係性を精査した．それぞれの推定式の関係性が明らかとなり，今後，英文書の性質に応じた各種リーダビリティ推定式の選択といった仕組みが望まれる．

③リーダビリティ推定式の公開

本研究で構築したリーダビリティ推定式は自由に使えるよう Web 上で公開した．今後，ローカル環境でも気軽に利用できるようにアプリケーションのような形でも公開していく予定である．

④リーダビリティ推定式の分析

今回構築したリーダビリティ推定式を精査し，学年に対するリーダビリティの推移を観察した．その結果，「中 1→中 2」と「中 3→高 1」で大きくリーダビリティが難化することが明らかとなった．

また各要因がリーダビリティにどのような影響を与えるか，といったことも示唆されたが，各要因に対する個別的な分析となっており，複数要因の総合的な解釈については今後の課題である．

(3) 学校文法の文法項目の検出のための基礎調査

リーダビリティ推定式に文書中（または文中単位）で使用されている学校文法の文法項目を組み入れることを検討した．そのための基礎的な研究として，学校文法に関する情報を付与した延べ約 1 万文の用例を電子化し，機械学習理論を応用した文法項目の自動検出ルールの記述を試みた．たとえば，仮定法では適合率 93.2%・再現率 73.3%の精度を得，その有効性と問題点を明らかにした．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 7 件）

①! 木村 恵，田中省作，八島 等，依田みずき：言語資源とその処理技術を活用したL2 語彙の習得レベル判定，英語コーパス研究，14 page (accepted).［査読有］

② 神谷健一，田中省作，北尾謙治：言語処理技術と教材作成の連携，自然言語処理，14 page (accepted).［査読有］

③ Kitao, K. and Tanaka, S.: Characteristics of Japanese Junior High School English Texts, 文化情報学, Vol.4,

No.1, pp.1-10 (2009). [査読有]

④ 神戸有美子, 田中省作: 英語圏のリーディング教材を用いた日本人英語学習者のリーダビリティ意識の調査, 統計数理研究所リポート, Vol.233, pp.49-59 (2009). [査読無]

⑤ 田中省作, 小山由紀江: 文分類モデルに基づいた多様なレベルのESP特徴表現抽出の試み, 統計数理研究所リポート, Vol.233, pp.21-33 (2009). [査読無]

⑥ 田中省作: 順序尺度としての学年を指標とした日本人英語学習者のための英文書リーダビリティの推定, 統計数理研究所リポート, Vol.233, pp.35-47 (2009). [査読無]

⑦ 小野 望, 田中省作, 持尾弘司: 母語学習者コーパスの基礎調査, 筑紫女学園大学人間文化研究所年報, Vol.18, pp.27-36 (2007). [査読有]

〔学会発表〕(計 10 件)
① 小林雄一郎, 田中省作, 後藤一章, 徳見道夫, 朝尾幸次郎: 文法情報の自動検出技術を用いたリーディング教材の作成と評価, 語彙研究フォーラム 2008 第1回JACETリーディング研究会・英語語彙研究会合同研究大会, 関西学院大学 (2008 年 12 月 6 日).

② 田中省作, 神戸有美子: 英語圏のリーディング教材を用いた日本人英語学習者のリーダビリティ意識の調査, 語彙研究フォーラム 2008 第 1 回JACETリーディング研究会・英語語彙研究会合同研究大会, 関西学院大学 (2008 年 12 月 6 日).

③ 田中省作: 最近の言語処理事情と言語教育への応用可能性, 筑紫女学園大学「コーパスを用いた日本語作文支援環境の構築」研究会, 筑紫女学園大学 (2008 年 12 月 1 日).

④ 小林雄一郎, 田中省作, 後藤一章, 徳見道夫, 朝尾幸次郎: 学校英文法コーパスの提案―デザインと応用可能性―, NLP若手の会第 3 回シンポジウム, 熱海金城館 (2008 年 9 月 23 日).

⑤ 田中省作: 言語処理を活用した英語教育事情―できそうなこと, できそうにないこと―, LEIS月例英語教育公開講座, 立命館大学 (2008 年 9 月 20 日).

⑥ 田中省作, 小林雄一郎, 徳見道夫, 朝尾幸次郎: 学校英文法コーパス構築の試み, 人工知能学会全国大会, ときわ市民ホール (2008 年 6 月 12 日).

⑦ Kitao, K. and Tanaka, S.: Authorized Junior High School English Textbooks in Japan -From the Viewpoint of Vocabulary and Readability-, GloCALL2007 , Hanoi University, Vietnam (2007 年 11 月 30 日).

⑧ 田中省作, 木村 恵, 八島 等, 依田みずき: 情報処理技術を活用した英語語彙の意味習得にかかわる要因の同定, 関東甲信越英語教育学会研究大会, 千葉商科大学(2007 月 8 月 16,17 日).

⑨ 田中省作, 富田康文: 順序尺度としての学年を指標とした英文書のリーダビリティ推定の試み, 外国語教育メディア学会, 名古屋学院大学(2007 年 8 月 9 日).

⑩ 北尾謙治, 田中省作: 中学校英語検定教科書の特徴―語彙とリーダビリティから―, 外国語教育メディア学会関西支部春季研究大会, 武庫川女子大学 (2007 年 5 月 12 日).

〔その他〕

(1)順序ロジットモデルに基づいた日本人英語学習者のためのリーダビリティ推定の Web ページ
http://www.cl.ritsumei.ac.jp/CALL/OLR/

６．研究組織

(1)研究代表者
田中　省作 (TANAKA SHOSAKU)
立命館大学・文学部・准教授
研究者番号：００３２５５４９

(2)研究分担者

(3)連携研究者