

「良い作文」に対する定量的分析

李在鎬・木戸光子・加藤あさぎ（筑波大学）・小浦方理恵（麗澤大学）

1. 背景と目的

本研究では、プレースメントテストでの応用を視野に入れながら、良い作文とは何かという問に対して、統計的な手法に基づいて考察する。具体的には、1) T大学の作文クラスにおいて2012春～2013秋の期間中（4クラス×4学期）に収集した作文を電子化した。2) コース担当者が当該レベルにおける達成度を主観判定した。3) 作文単位でテキスト情報量を計算した。4) 分散分析と決定木分析を行い、達成度を説明できる言語的要因は何であるかを分析した。分析の結果、レベルによって達成度に関する言語的要因が異なっていることが明らかになった。本研究の調査結果は主観的になりがちな達成度の判別に関する科学的な検証であり、「良い作文」が持つ多様性、複雑性の一端を明らかにするものと考えられるため、文章の自動評価の基礎研究になる。

2. 先行研究と問題提起

良い文章、良い作文とは何かという問に対しては、いわゆる評価基準を提案する先行研究があるが、パフォーマンス評価に共通する信頼性・妥当性の問題に対しては、根本的な問題が残る形になっている（作文評価については、田中・阿部 2014 参照，パフォーマンス評価の問題点については李（編著）2015 参照）。

従来の作文評価の多くは、パフォーマンス評価に見られる一般的課題に対して、根本的には解決できていないことが多い上に、評価指標の多くが経験的な要素で決まるため、データに基づく妥当性の検証が難しい。この課題を解決するためには、評価指標に対して反証可能な形で議論を構成する必要があると考える。とりわけ、評価者に依存しない客観的評価指標を抽出することと評価指標間の階層関係を明確にすることが重要と考える。この課題を踏まえ、本研究では、プレースメントテストでの利用を前提に、機械的な処理モデルを用いて、いつ、誰が測定してもゆれのない評価軸を構築するとともに、統計的なモデリングを行うことで、妥当性・汎用性についても明らかにする。研究調査の具体的な課題として、以下の2点を設定する。

- 1) 良い作文と悪い作文を仕分ける言語的基準を抽出する。
- 2) 全集団において良い作文と悪い作文を仕分ける要因（の階層構造）を明らかにする。

1) は到達度評価 (achievement assessment) に関わる問題提起である。2) は熟達度評価 (proficiency assessment) に関わる問題提起である。1) の課題設定をとおして、作文の良し悪しに関わる判断に、習熟度がどのようなバイアスを与えるのか明らかにしたい。2) の課題設定をとおして、文章の読み手にとって良い作文が持つ普遍的な要因を明らかにするとともに、その要因の強弱をデータサイエンスの観点から明らかにする。

3. データと分析方法

分析に用いたデータは、2012年春学期から2013年秋学期において4つのレベル（中級の下(J500)、中級の上(J600)、上級の下(J700)、上級の上(J800)）の受講者が手書きで作成した作文である（作文の総数は582件、総語数は106,880語、一作文の平均字数は280.7字）。このデータに対して、以下の手順で分析を行った。

- ステップ1: 手書き作文の電子化
- ステップ2: コースの担当教員による達成度評価
- ステップ3: 該当テキストの言語処理と数量化
- ステップ4: 統計分析

ステップ1として、手書きの作文を手入力し、電子化した。ステップ2として、各コースを担当した教員が、全作文に目をとおり、当該レベルとして、非常によくかけている作文（A判定作文）と当該レベルとして、よくない作文（C判定作文）をそれぞれ10編ずつ選んだ。非常によくかけている作文の判断基準は、担当教員の主観的判断で行った。本研究は、教員の主観的判断の中身を計算モデルで再構築し、内部構造を明らかにすることを目的としている。なお、分析では、A判定とC判定のもの以外に、どちらにも属さない作文（＝良いとも悪いとも言えない作文）をB判定作文にし、ほかと同様に10編を選んだ。A判定は、よくできた作文、B判定は、まあまあできた作文、C判定は、悪い作文という想定のもとで分析を行った。

中級の下レベル（J500）から上級の上レベル（J800レベル）の4レベルで、A～C判定の各10編が集まり、合計で120編（4レベル×3水準×10編＝120編）の分析データが得られた。これをステップ3で<http://jreadability.net/>を使い、数量データに変換した（図1）。

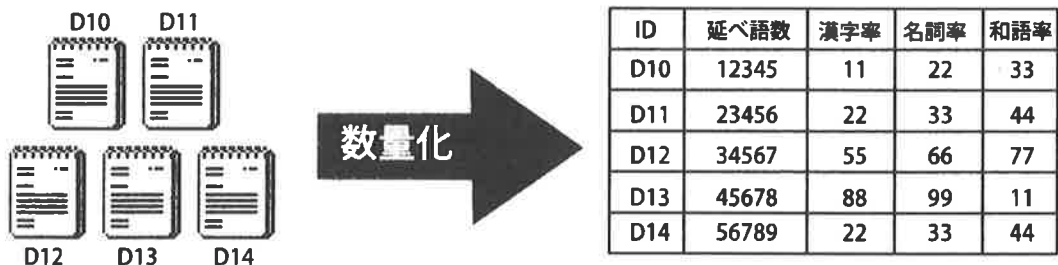


図1. 数量化のイメージ

図1に示すとおり、120編の文章データを延べ語数や異なり語数、さらには動詞や名詞類の使用頻度、含有語彙レベルを集計した。数量化に使用した変数は、全部で18個あり、語種別の使用率、語の長さ、文の長さ、品詞の使用率、語彙難易度の使用率などが入っている。ステップ4では、IBM SPSS ver.23を使用し、達成度スケールを従属変数、数量化に用いた変数を独立変数に指定し、分散分析と決定木分析を行った。

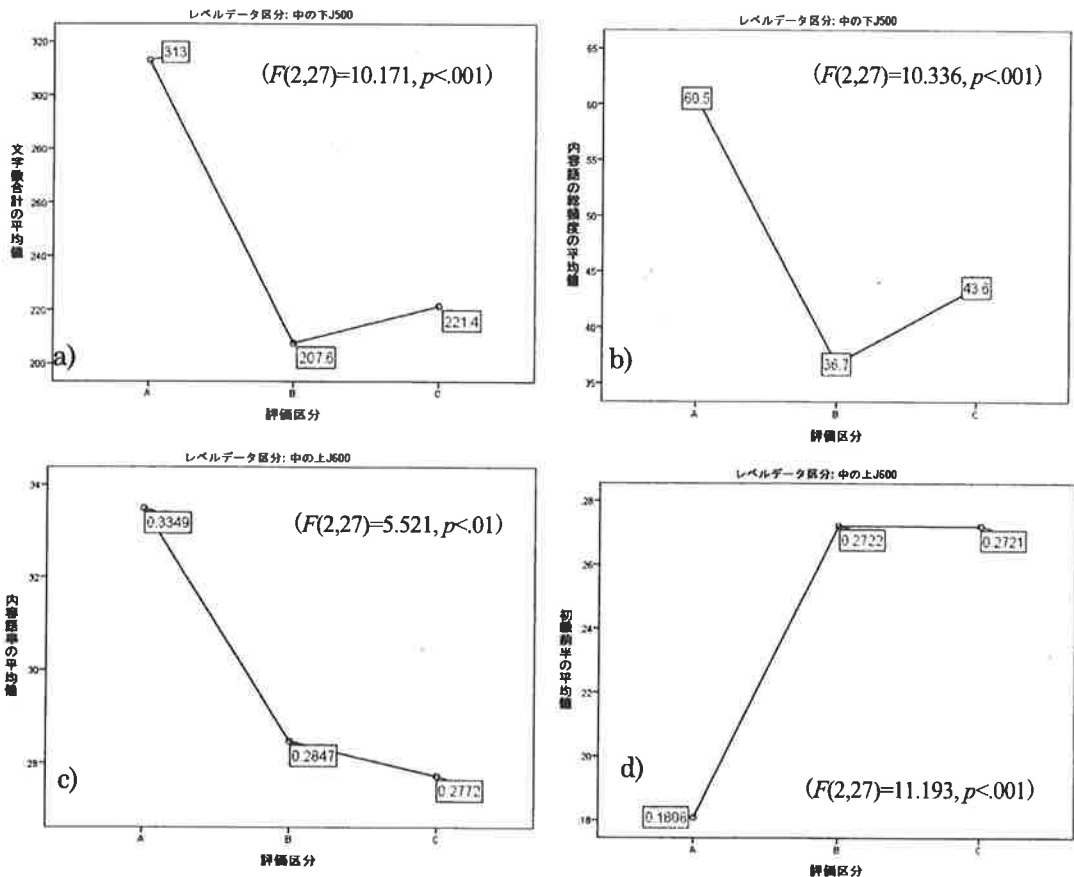
4. 結果と考察

統計分析の結果を述べる。まず、分散分析の結果を述べてから、決定木分析の結果を述べる。分散分析の結果については、従属変数に対して統計的な有意差が確認された項目のみを取り上げることにする。

まず、「中級の下」レベルでは、合計文字数や内容語の総頻度において、習熟度による産出頻度の統計的有意差が確認された（図 2a, 図 2b）。これは、主に作品全体の長さによる差として説明できる。つまり、このレベルにおいては、ある程度の長さを持ち、内容語の絶対的な出現頻度が作文の良し悪しに影響を与えていると考えられる。

次に「中級の上」レベルでは、初級前半語彙や内容語の使用率において有意差が確認された（図 2c, 図 2d）。これは、語彙の質的な属性による差として説明できる。とりわけ、内容語の相対的な使用率が高いほど良い作文であると判定されている。また、語彙難易度の部分では、初級レベルの語彙が多いほど、C 判定もしくは B 判定になっており、ある程度、難しい語彙が使用されているかどうかの評価に影響していると考えられる（語彙レベルの判定は「日本語教育語彙表 (<http://jhlee.sakura.ne.jp/JEV.html>)」に準拠して行った)。

次に、「上級の下」レベルでは 1 単語の平均的な長さにおいて統計的有意差が確認された（図 2e）。これは、語彙知識として長い語彙を知っているかどうかの差として説明できる。



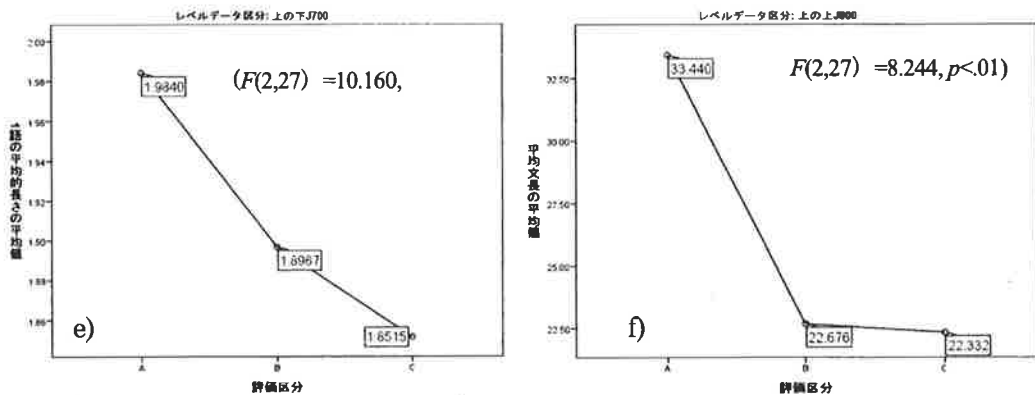


図2. 分散分析の結果

最後に「上級の上」レベルでは、一文の平均的な長さにおいて、習熟度による産出頻度の統計的有意差が確認された(図2f)。これは、複雑な文章作成の能力の差として説明できる。つまり、複文などのように一文の物理的な長さが重要な要因として関与しているのである。

図2の結果から得られる示唆として、日本語教師にとって良い作文という判断は、学習者の習熟度によってかなり異なるという事実が明らかになった。全体的な傾向として、中級では語彙によって表現される内容面の豊かさや作文全体の長さが重要な要因になっている。一方、上級では語や文章の長さが良い作文を線引きする要因になっており、どの程度、正確に表現することができるかという観点が重要になっていると考えられる。

次に、習熟度の問題を超えて、読み手にとって良い作文とは何かという問題を考えてみたい。すべての学習者における共通属性を捉えるべく、2つの決定木分析を行った。1つ目は、A～C判定の到達度を従属変数に、2つ目は、J500～J800の習熟度を従属変数にして、分析してみた。図3は到達度に対する決定木分析の結果であり、図4は、習熟度に対する決定木分析の結果である。なお、決定木分析とは、従属変数に対して独立変数の統計的な有意差をもとに変数を階層化しながらデータを分類する機械学習のアルゴリズムの一つである。本研究では、IBM SPSS ver.23のClassification Treeを使い、決定木分析を行った。

図3の決定木分析の結果については、以下のことが言える。

1. 1語の平均的な長さや内容語の総頻度と作文全体の延べ語数で全体のデータが分類されている。
2. データの分類にとって1語の平均的な長さをもっとも重要である。特にA判定とB・C判定の特徴抽出にとって1語の平均的な長さをもっとも大きな要因になっている。

この結果をもとに、A判定を受けた全作文の共通要因とC判定を受けた全作文の共通要因を次のように推定することができる。1) A判定を受けた作文に共通する性質として作文全体における1語の平均的な長さが1.9拍以上、作文全体において内容語が63語以上、使用されている。2) C判定を受けた作文に共通する性質として作文全体における1語の平均的な長さが1.9拍未満、作文全体において内容語が63語未満であり、全体の語数が168語未満である。

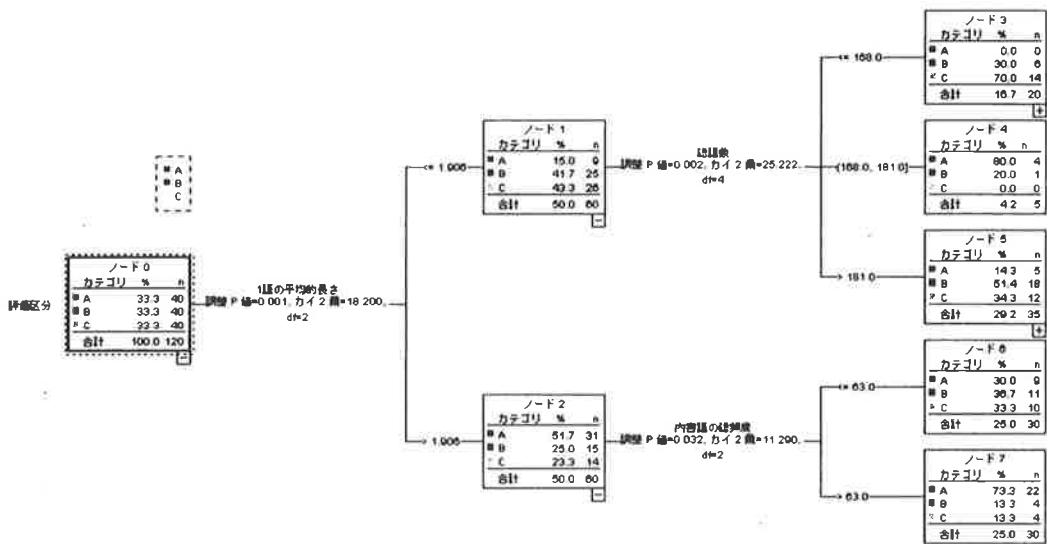


図3. 達成度に対する決定木

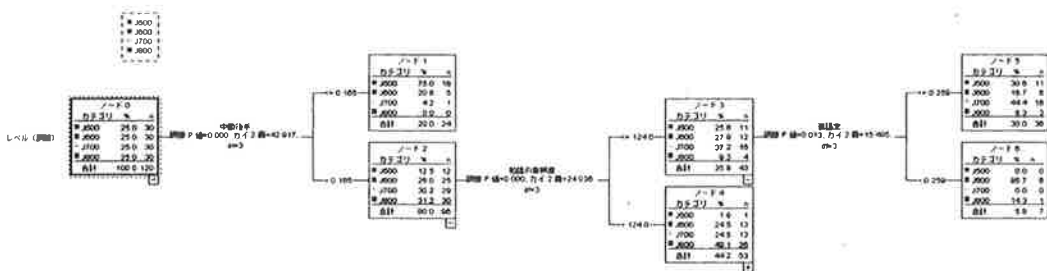


図4. 習熟度に対する決定木

次に図4の習熟度に対する決定木分析の結果については、以下のことが言える。

1. 中級後半の語彙が使えるかどうかで、中級の下(J500)とそれ以外のグループが別れる。
2. 和語の総頻度で上級の下とそれ以外のグループが別れる。
3. 漢語の使用率によって、中級の上(J600)と上級の下(J700)が別れる。

上述の特徴をもとに、図3に対して行ったのと同じように、中級の下(J500)～上級の上(J800)のレベルに共通する性質を推定できる。例えば、中級の下(J500)に共通する特徴としては次のように記述できる。テキスト全体における中級後半語彙の使用率が16%未満であり、和語の総頻度が124語未満、漢語率が25%未満である。

5. まとめと課題

本研究では、良い作文とはなにかという問に対して、T 大学で収集した大規模な作文データを、計算モデルをつかって定量的に分析した。本研究が設定した2つの課題に関して明らかになった点を述べる。

課題1) 習熟度集団別に、良い作文と悪い作文を仕分ける言語的基準を抽出する。

→分散分析で4つの習熟度別に、良い作文と悪い作文の言語情報量を明らかにした。中級の下(J500)では語数や内容語の総使用量が、中級の上(J600)では内容語の相対的な使用頻度、上級の下(J700)では1語の平均的な長さ、上級の上(J800)では1文の平均的な長さにおいて有意差が確認された。

課題2) 全集団において良い作文と悪い作文を仕分ける要因(の階層構造)を明らかにする。

→決定木分析で到達度別・習熟度別にサブクラスを階層化した。到達度に関わる要因としては、1語の平均的な長さや内容語の総頻度などが重要であることが明らかになった。習熟度に係る要因としては、中級語彙や和語や漢語の使用頻度が重要であることが明らかになった。

本研究では人の判定を計算モデルで考察したが、プレースメントテストに実装するための今後の課題としては、計算モデルによって分類したデータを人の判定で検証する作業が必要である。

*謝辞：本研究は、2013年度および2014年度の筑波大学・学群教育充実事業経費の補助を受けて行いました。データ入力および判定においては、本発表の第2著者、第3著者、第4著者のほかに、筑波大学グローバル・コミュニケーション教育センター(旧留学生センター)の「書く」クラス担当の石上綾子氏、田中孝始氏、長戸三成子氏、山本千波氏、渡邊芙裕美氏に大変お世話になりました。深くお礼申し上げます。

【参考資料】

- 田中真理・阿部新(2014)『Good writing へのパスポート—読み手と構成を考えた日本語ライティング』くろしお出版
- 李在鎬(編著)(2015)『日本語教育のための言語テストガイドブック』くろしお出版。