英語学習者のアウトプットにおける語彙の多様性研究の現在と今後の課題

Current trends and issues of lexical diversity studies in productive texts of English learners

小島ますみ
Masumi　KOJIMA

Abstract

The purpose of this study is to examine the reliability and validity of widely used lexical diversity (LD) measures including type/token ratio (TTR), standardized TTR, Guiraud Index, Herdan Index, D, HD-D, and Measure of Textual Lexical Diversity (MTLD) through a survey of previous studies and discussion. LD refers to the range of vocabulary within a text and avoidance of repetition; it is often considered to reflect the vocabulary proficiency or productive lexical richness of a learner. A serious problem with LD measures is that they are dependent on text length, which indicates a lack of reliability. Even when LD measures overcome the problem of text length, researchers need to carefully examine their validity as developmental measures of the productive vocabulary of second language (L2) learners. This paper also discusses the limitation of LD measures and suggests directions for further studies.

Keywords :語彙習得，発表語彙，語彙の多様性

1. はじめに

これまでの第二言語における語彙習得研究の多くは, 学習者の受容語彙知識を研究対象としてきた。しかしながら Nation (2007) は，学習者の持つ語彙知識の全体像をつかむためには, 学習者によって語彙がどのように使用されるかを調べる必要があると述べている。近年のコンピュータや学習者コーパスの発達に伴い, 学習者が実際に産出する語彙を分析・評価する試みが盛んになってきた。Laufer (1998) は，第二言語学習者が母語話者と顕著に異なるのは, スピーキングやライティング時の発表語彙においてであると述べている。つまり，母語話者が多様で広範な語彙を使用するのに対し, 学習者は限られた語彙しか使用しない。また学習者間でも，習熟度が上がるにつれて幅広い語彙が使用されるようになると指摘している。

語彙の多様性とは，書き手や話し手が，語の繰り返しを避けさまざまに異なる語を使用する程度と定義される（Malvern, Richards, Chipere & Duran, 2004）発表語彙の発達した学習者は, 語の繰り返しを避けさまざまな言い換え表現を使用するなど, より多様な語彙を産出すると考えられる。このため言語習得研究では，語彙の多様性はメンタルレキシコン内の語彙サイズやその語彙知識を効果的に使用する能力を表していると捉えることが多い。これまで提案された語彙の多様性を量的に示す指標は, テクストにおける異語数と総語数に基づき算出されるものである。このような語彙の多様性指標は，母語習得，第二言語習得，医学臨床等の分野で, 一般的な言語の発達指標として広く用いられてきた (Malvern & Richards, 1997)。

第二言語習得研究において，語彙の多様性指標は主に 2 種類の研究で応用されている。1 つ目は，ライティングやスピーキングの質に影響する要因の研究であり，2 つ目は, 語彙知識と語彙使用の関係についての研究である。ライティングの質に影響する要因の研究としては, Arnaud (1984), Engber (1993), Linnarud (1986) などがあげられる。スピーキングの質に影響する要因の研究も近年盛んになっており，Richards & Malvern (2000)，Vermeer (2000)，Malvern & Richards (2002), Koizumi (2005) らが, そのような研究課題に取り組んでいる。これらの研究では，学習者の言語産出における語彙の多様性や総語数，平均文長，統語論的特長などを調べ, ライティングやスピーキングの評価と最も相関の強い指標を調べたり, 重回帰分析により評価を予測する回帰式を作成するなどして, ライティングやスピーキングの質に影響する要因のモデル化を行おうとするものである。語彙知識と語彙の関係についての研究は，Vermeer (2000)，石川 (2008) などがあげられる。これらの研究で

は, 学習者のライティングやスピーキングにおける語彙の多様さと語彙テストの結果を比較し, 両者の相関関係を調べたり, それらの相関関係が学習者の熟達度の変化に伴いどのように変化するかを調べるものである。石川 (2008) では構造方程式モデリングを使用し, 学習者の語彙知識と語彙使用についてモデル化を行う試みがなされている。

語彙の多様性指標の代表は TTR であるが，テクストの長さの影響を受けるという大きな問題があり, この問題を解決するためのさまざまな試みが 70 年以上もの間続けられてきた (e.g. Guiraud, 1954; Johnson, 1944; Herdan, 1960; Malvern & Richards, 1997; McCarthy & Jarvis, 2010)。これらの指標を応用する研究者は, それぞれの指標の仕組みや特徴を知り, 指標の信頼性や妥当性について十分調査する必要があるが, そのような取り組みはほとんどなされていないのが現状である。テストや指標の信頼性とは，スコアの一貫性・安定性の程度と定義され，妥当性とは，測ろうと意図しているものを測定できている度合いと定義される (Brown, 1996)。学習者が持つ「語彙の多様性」という特性を測りたい場合, そのような特性はテクストの長さによって変化するとは考えられないため, テクストの長さにかかわらず一貫して安定したスコアを返すことが期待される。また, 学習者の語彙発達や言語発達を表すものとして語彙の多様性指標のスコアを利用したいのであれば, そのような前提は妥当かどうか検討する必要がある。

本研究は，代表的な語彙の多様性指標として，TTR，テクストの長さをそろえた TTR，標準化 TTR, Guiraud Index, Herdan Index, D, HD-D, MTLD を取り上げ，それぞれの仕組みと特徴を概観し, テクストの長さの問題をどの程度克服できているのかという信頼性の問題や, 学習者の語彙発達や言語発達指標としての妥当性の問題について検討し，今後必要な語彙の多様性研究についての提案を行う。

2. さまざまな語彙の多様性指標

2.1 TTR (Type-Token Ratio)

語彙の多様性指標の中で，最もよく知られ，さまざまな研究で使用されているのは，TTR (type-token ratio) である。TTR は次の式で算出される。

$$\frac{\text{異語数（type）}}{\text{総語数（token）}}$$

例えば，We will continue working as long as we have daylight. という英文があったとする。この文に含まれる総語数は 10 であるが，we と as は 2 回ずつ使用されているため，異語数は 8 となる。そこで，このセンテンスにおける TTR は 0.8 と算出される。発表語彙の発達した学習者は，語の繰り返しを避け，さまざまな言い換え表現を使用するなど，より多様な語彙を産出すると考えられるため，TTR は高くなると予想される。

しかしながら，TTR はさまざまな言語発達指標と負の相関を持つことが, 多くの研究で報告されている。例えば, Vermeer (2000) は，オランダ語を母語または第二言語とする 70 人の子どもを対象に，2 種類の語彙テストと発話データにおける TTR との相関を調べたところ，ほぼ無相関であった (r =-.13, -.09)。Chipere, Malvern & Richards (2004) では, 英語を母語とする子どもが書いた 918 の物語文を対象に，総語数やスペリングの正しさと TTR の相関を調べたところ，弱い負の相関が観察された(r =-.26, -.14)。

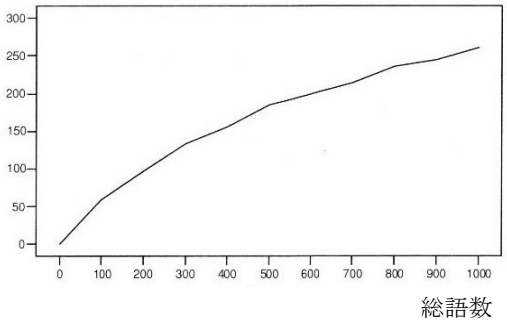

図 1　『創世記』における総語数と異語数の変化
(van Hout & Vermeer, 2007, p. 96, Figure 1)

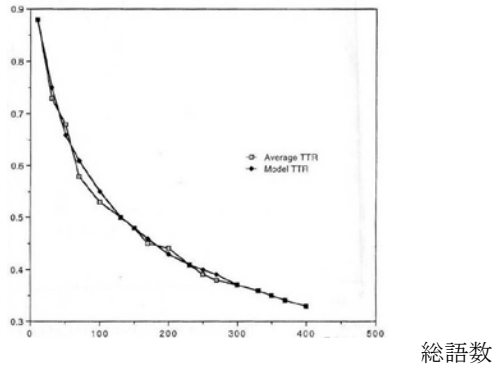

図 2　発話データにおける総語数と TTR の変化
(Malvern & Richards, 1997, p. 66, Figure 2)

言語の発達とともに TTR が減少するという現象は，テクストの長さを統制することなく TTR が使用された場合に観察されている。総語数は，テクストの長さが長くなるとともに増加するが，異語数の場合は，テクストの長さが長くなるほど，新出語が使用される確率は低くなり，増加幅は減少する (図 1 参照)。したがって異語数÷総語数で定義される TTR は，テクストの長さが長くなるとともに減少するという性質を持つ (図 2 参照)。

母語習得研究，第二言語習得研究ともに，熟達度の高い学習者ほど，産出量 (総語数) が多いという報告は多数ある (e.g. Chipere, Malvern & Richards, 2004; Daller & Xue, 2007; Vermeer, 2000)。このような総語数と負の相関を持つ TTR は, 言語の発達指標として適切でないことは明らかである。

TTR がテクストの長さの影響を受け，テクストの語数が増加するとともに減少するという問題を解決するため，さまざまな方法が提案されてきた。それらの解決策は大きく以下の 4 種類に分類することができる

1. テクストの長さを揃え，TTR を測定
2. TTR の式を変換
3. 数学的に TTR 曲線を予測
4. 主題に関する飽和の概念を利用

1 の例として，テクストの長さを最短データにそろえて測定した TTR や，標準化 TTR があげられる。2 の例として，Guiraud Index (Guiraud, 1954) や Herdan Index (Herdan, 1960) があげられる。3 の例として, D (Malvern & Richards, 1997) や HD-D (McCarthy & Jarvis, 2007) があげられる。4 の例として，Measure of Textual Lexical Diversity (McCarthy & Jarvis, 2010) があげられる。2.2 節以降では，これらの指標の特徴や問題点について, これまでの研究結果をまとめる。

2.2 テクストの長さを揃えた TTR

TTR がテクストの長さの影響を受け，テクストの語数が増加するとともに減少するという問題を解決するため，さまざまな方法が提案されてきた。最も単純な方法は，テクストの長さを揃え, TTR を測定するという方法である。例えば Laufer (1991) は，収集した作文データの最短テクストが 250 語だったことから, すべてのテクストにおいて, 最初の 250 語を分析対象とした。Arnaud (1984) は，収集した作文データの最短テクストが 180 語だったことから，すべてのテクストにおいて, コンピュータにランダムに選ばせた 180 語を分析対象とした。

しかしながら, このようなテクストの長さを揃えた TTR の問題点も指摘されている。まず，すべてのデータを活用できないという問題がある。学習者の産出データにおける総語数というのは，かなりのばらつきがある。すべてのデータを最短テクストに揃えると, データのごく一部しか活用できないことになる。学習者は語彙をランダムに使用するわけではなく, 結論の部分に重要な語彙を使用するなど偏りがあると考えられるため, 学習者の産出語彙全体を評価する方が適切である。また，Chipere, Malvern & Richards (2002) は，テクストの長さを何語でそろえるかというカットオフポイントが研究者によって異なるため, 研究間の比較が困難であるという問題を指摘している。

また，短いテクストの場合，TTR の信頼性が低いという問題もある。Hess, Haug & Landry (1989) は，52 人の英語を母語とする子どもから収集した最短 400 語の発話データを 50 語×8, 100 語×4, 200 語×2 に分け，それぞれのセグメントにおける TTR を求め，相関を調べることで TTR の信頼性を調査した。結果は有意な相関はみられず, Hess らは 200 語以下の短いテクストにおける TTR の信頼性は低いと結論づけている。また Hess, Sefton & Landry (1986) は, TTR で信頼性係数.70 を確保するためには, 350 語以上，信頼性係数 .80 では，550 語以上必要であると報告している。これらの数値は，スピアマン・ブラウンの予言公式を利用することで算出された。まず彼らは，83 人の英語を母語とする子どもから収集した平均 250 語の発話データを 50 語×4, 100 語×2 に分け, それぞれのセグメントにおける TTR の相関係数を求めた。これらの相関係数を基に，スピアマン・ブラウンの予言公式を利用し，信頼性係数.70 と.80 を得るために必要な総語数が算出された。Hess らは信頼性係数 .70 について，「最低限必要な信頼性」と述べている。先ほどの Laufer (1991) や Arnaud (1984) の例では，この総語数の基準に達しておらず，信頼性の低いデータを使用した可能性が高い。また，日本の英語教育の現状からして, すべての被験者から 350 語以上のデータを収集するというのは，大変困難である。特に初級の学習者が対象の場合，非現実的であると言える。

また Vermeer (2000) は，テクストの長さを揃えた場合，１つのトピックでたくさん産出する学習者は, トピックの変化が多い学習者に比べ, 同じ語数当たりのトピック数が少なくなり, 語彙の多様性が減少することを指摘している。このことから Vermeer (2000) は, 安易にテクストの長さを揃えるべきではないと主張している。例えば，同じ 200 語の発話でも, 家族のトピックで母親の話だけをする学習者と，母親・父親・兄弟の話をする学習者では，後者の語彙の多様性の方が高いことが予想される。

2.3 標準化 TTR (standardized TTR)

Johnson (1944) は，TTR がテクストの長さの影響を受けることを指摘し，テクストの長さの影響を抑える方法の 1 つとして，標準化 TTR を提案している。標準化 TTR は，テクストを一定の短いセグメントに分割し, 各セグメントにおける TTR を求め，その平均値で表される。テクストの長さを揃えた TTR と比較した利点は，より多くのデータを活用することができるという点である。例えば，200

語のデータと 300 語のデータがあった場合，100 語ずつのセグメントに区切れば，すべてのデータを使用できる。セグメントの長さは，研究によりさまざまに異なっている。例えば Richards & Malvern (2000) は，教師の発話については 100 語ずつに分割して標準化 TTR を算出したのに対し，生徒の発話については，30 語ずつに分割して標準化 TTR を算出している。

テクストの長さを揃えた TTR に比べれば，より多くのデータを活用することができるが，さまざまな長さのテクストを産出する学習者集団に対し，すべてのデータを利用することはやはり困難である。例えば，100 語ずつのセグメントに区切って標準化 TTR を算出する場合，220 語など 100 語で割り切れないテクストの場合は，端数切り捨てとなる。

Malvern & Richards (2002) は，標準化をするためのセグメントの長さが研究者によって異なるため，研究間の比較が困難であるという問題を指摘している。TTR が，総語数が増加するとともに減少する傾向を持つのと同様に，通常短いセグメントに基づく標準化 TTR の方が，より長いセグメントに基づく標準化 TTR よりも，スコアが高い傾向となる。このため，異なる長さのセグメントに基づく標準化 TTR を直接比較することができない。Malvern らは，学習者の産出データにおける総語数と異語数の関係というのはダイナミックなものであるのに対し，標準化 TTR は，そのダイナミックな TTR 曲線の 1 点しかとらえていないと批判している。

Malvern らはまた，標準化 TTR はセグメント内での語彙の多様性を測定することになり，セグメントを越えて語彙が繰り返し使用される場合は考慮しないため，適切な結果が得られない点を指摘している。彼らはまた，Richards & Malvern (2000) の結果より，標準化 TTR は，発話量や平均文長などの他の発達指標に比べ個人差があまりなく，分散が小さいという問題も指摘している。

2.4 Guiraud Index

Guiraud Index は，TTR の下降曲線を直線に近づけるために，Guiraud (1954) により提案された変換方法であり，次の式で定義される。

$$\frac{\text{異語数}}{\sqrt{\text{総語数}}}$$

Guiraud index は TTR よりも，他の言語発達指標と相関が高く，明らかに熟達度の異なる 2 つのグループをより良く区別することが，多くの研究で実証されている。例えば Vermeer (2000) は，オランダ語を母語または第二言語とする 70 人の子どもを対象に，2 種類の語彙テストと発話データにおける Guiraud Index の相関を調べたところ，有意な正の相関が観察された (r = .46, .48)。また，オランダ語を母語とする子どもと，第二言語とする子どもを比較したところ，発話データにおける Guiraud Index に有意差がみられた。この結果は，同じデータを TTR で分析した結果と比較すると対照的である。すなわち，TTR と 2 種類の語彙テストの相関はほとんどなく (r =-.13, -.09)，オランダ語を第二言語とする子どもの TTR の方が，オランダ語を母語とする子どもの TTR よりも高い結果となっている。

また，Daller & Xue (2007) は，外国語として英語を学習する中国人大学生グループとイギリスの大学で勉強している中国人留学生グループ計 50 人を対象に，漫画の絵を見て口頭で描写するタスク (picture description task) を使用し，収集した発話データにおける Guiraud Index や TTR などの指標を比較したところ，TTR では有意差がなかったのに対し，Guiraud Index では有意差が見られ，イギリスの大学に留学中の中国人の Guiraud Index の方が，外国語として英語を学習する中国人大学生の Guiraud Index よりも有意に高い結果となった。

Van Hout & Vermeer (2007) は， オランダ語を母語または第二言語とする 32 人の子どもを対象に，2 種類の語彙テストを行い，発話データを収集した。発話データにおける Guiraud Index は，2 種類の語彙テスト，異語数，総語数，レマの数と有意な正相関があった (それぞれ，r =.40, .49, .81, .54, .79)。これに対し，TTR と 2 種類の語彙テストには有意な相関がなく (それぞれ r = .02, -.04)，TTR と異語数，総語数，レマの数との相関は，有意な負の相関であった (それぞれ，r =-.63, -.84, -.63)。

Guiraud Index は TTR に比べ，よりよい発達指標であると言えそうだが，TTR の下降曲線を一定な直線に変換するわけではない。理想的には図 3 のように，1 人の産出データにおいて，TTR の変換値はテクスト中のどの長さを取っても，一定になることが望まれる 。

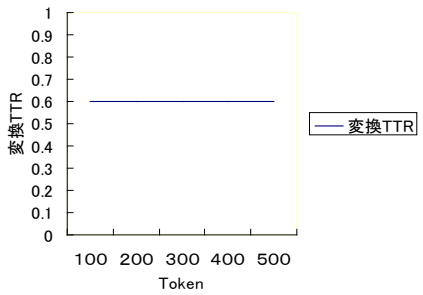

図３ 理想的な変換後の TTR

Guiraud Index で変換した TTR 曲線は，図 3 のような一定な直線になるわけではなく, 逆に上昇曲線に変換されることが実証されている。Hess, Haug & Landry (1989) は，52 人の英語を母語とする子どもから収集した最短 400 語の発話データを 50 語×8, 100 語×4, 200 語×2 に分け，それぞれの語数における Guiraud Index の平均値を求めた。分散分析と多重比較の結果, すべての語数グループ間で有意差あり, 同じ子どものデータでもテクストの長さが長いほど，Guiraud index の値は大きくなることを証明した。

McKee, Malvern & Richards (2000) らも指摘するように, Guiraud Index は総語数の関数であるため, テクストの長さの影響を受けるのは必須であると考えられる。

Daller & Xue (2007) は, Guiraud Index と総語数の相関が高いのは，より熟達度の高い学習者はより産出量が多く，Guiraud Index も高くなる結果であり，必ずしも Guiraud Index がテクストの長さの影響を受けるとは言えないと述べている。しかし彼らは, 同一被験者の Guiraud Index が, 語数の増加に伴いどのように変化するかは調べておらず, Guiraud Index とテクストの長さの関係について, 十分な調査を行っていない。

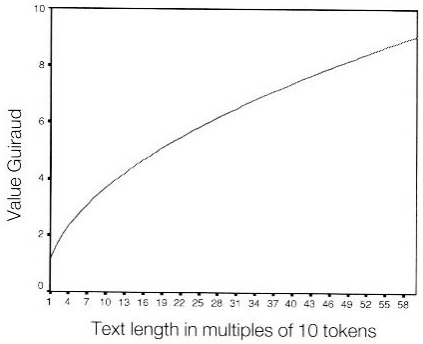

図４　テクストの長さと Guiraud Index との関係
(Daller & Xue, 2007, p. 161, Figure 3)

なぜ Guiraud Index は, 他の言語発達指標と相関が高く，TTR に比べ熟達度の異なる 2 つのグループをより良く区別するのだろうか。Daller & Xue (2007) は，6 つの指標と総語数の相関係数を調べたところ, Guiraud Index が最も高く，r =.64 であったと報告している (図 4 参照)。この場合，Guiraud Index の 41%は総語数が説明するという計算になる。Vermeer (2000) らの研究で，Guiraud Index と相関が高いと報告された言語発達指標は，総語数，異語数，レマの数など, テクストの長さが長くなるとともに増加する指標が主であった。テクストの長さと相関の高い指標は，Guiraud index とも相関が高いという結果に過ぎないと考えられる。

また, 2.1 節で述べたように, 熟達度の高い学習者ほど，産出量が多い (テクストの長さが長い) という報告は多数ある。Guiraud Index が増加した場合，熟達度の向上により産出量が増えたことによる可能性があり, 必ずしも語彙の多様性が向上したとは言えないと考えられる。Guiraud Index は語彙の多様性のみを測っているのではないため，語彙の多様性指標として，適切なものではない。

それでは，テクストの長さを揃えて Guiraud Index を使用すれば，問題は解決されるだろうか。テクストの長さの影響を抑えるために TTR を Guiraud Index に変換しているのであるから，テクストの長さを揃えて Guiraud Index を使用する意義は薄い。2.2 節や 2.3 節議論した，テクストの長さを揃えた TTR や標準化 TTR の問題点がそのまま当てはまる。Hess, Haug & Landry (1989) は，Guiraud Index でも TTR と同様に，テクストが短い場合，信頼性が低いことを実証している。TTR と同様に，Guiraud Index で信頼性係数.70 を確保するためには，350 語以上必要である (Hess, Sefton & Landry, 1986)。

Guiraud Index とよく似た指標に，修正 TTR (corrected TTR) と呼ばれるものがある。これは Carroll (1964) により提案されており, Guiraud Index に $1/\sqrt{2}$ をかけたものである。しかし修正 TTR は, Guiraud Index と完全に相関し (r = 1.0)，Guiraud Index と同様に，語数の増加とともに増加する (Hess, Haug & Landry, 1989)。修正 TTR で Guiraud Index の問題はまったく解決されない。

2.5　Herdan Index

Herdan Index は，TTR の下降曲線を直線に近づけるために, Herdan (1960) により提案された変換方法であり, TTR の分子・分母を対数変換するものである。Herdan Index は，次の式で定義される。

$$\frac{Ln\ \text{異語数（type）}}{Ln\ \text{総語数（token）}}$$

Herdan Index と言語発達指標の関係を調べた研究がいくつかあるが，Herdan Index が TTR と比較し，よりよい発達指標であるとは言いがたい結果となっている。例えば, Vermeer (2000) は，オランダ語を母語または第二言語とする 70 人の子どもを対象に，2 種類の語彙テストと発話データにおける Herdan Index との相関を調べたところ, ほぼ無相関であったと報告している (r = -.06, -.07)。また，オランダ語を母語とする子どもと, オランダ語を第二言語とする子どもの Herdan Index を比較したところ, 有意差はみられなかった。TTR でも似たような結果が見られたことから，Herdan Index と TTR は似たような結果を返すと考

えられる。

Herdan Index は，TTR の下降曲線をどのような曲線または直線に変換するのであろうか。理想的には図 3 のように，１人の産出データにおいて，変換後の TTR はテクストの長さが異なっても，一定になることが望まれる。Hess, Haug & Landry (1989) らの報告によって，TTR を Herdan Index で変換しても，依然として下降曲線であることが実証されている。Hess らは，52 人の英語を母語とする子どもから収集した最短 400 語の発話データを 50 語×8, 100 語×4, 200 語×2 に分け，それぞれの語数における Herdan Index の平均値を求めた。分散分析と多重比較の結果，すべての語数グループ間で有意差あり, 同一人物のデータでも総語数が多いほど，Herdan Index の値は小さくなることを証明した。

テクストの長さを揃えて Herdan Index を使用しても, 問題は解決されない。テクストの長さの影響を抑えるために TTR を Herdan Index に変換しているのであるから，テクストの長さを揃えて Herdan Index を使用する意義は薄く，2.2 節や 2.3 節で議論した，テクストの長さを揃えた TTR や標準化 TTR の問題点がそのまま当てはまることになる。Hess, Haug & Landry (1989) は，Herdan Index でも TTR と同様に，テクストが短い場合，信頼性が低いことを実証している。TTR と同様に，Herdan Index で信頼性係数.70 を確保するためには，350 語以上必要である (Hess, Sefton & Landry, 1986)。

2.6 D

Malvern & Richards (1997) は，TTR 曲線全体を数学的に予測することで，TTR がテクストの長さの影響を受け，語数が増加するとともに減少するという問題を解決する方法を提案した。彼らは Sichel (1986) が発表した TTR 曲線を表す関数式を単純化し，学習者の TTR 曲線は，下記の関数式で表せるとした。

$$TTR = \frac{2}{DN}\left[(1+DN)^{\frac{1}{2}} - 1 \right]$$

この関数式は，縦軸を TTR とし，横軸を語数 (N) とする TTR 曲線を表している。D は TTR 曲線全体がどこに位置するかを決定するパラメータである。上記関数式をモデルとし，コンピュータプログラムによりパラメータ D を調整することで，学習者のテクストから得られた TTR データに最も適合する TTR 曲線を推測する。語彙が多様であるほど，TTR 曲線がグラフ上部に位置し，それに伴いパラメータ D の値が増加することから，この D を指標とする (図 5 参照)。

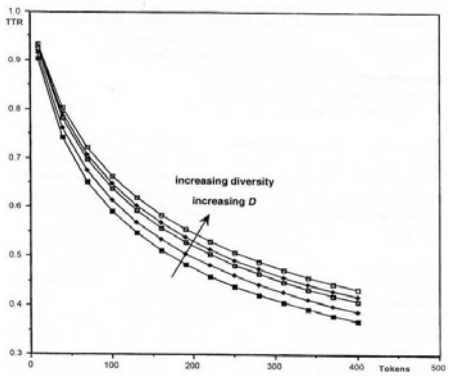

図５　TTR 曲線と D 値の関係

(McKee, Malvern & Richards, 2000, p. 325, Figure 1)

CHILDES (Child Language Data Exchange System) プロジェクトの CLAN (Computerised Language Analysis) プログラムの中に，パラメータ D を算出する vocd プログラムが搭載されている。CHILDES プロジェクトは，言語習得研究を促進するために，子どもの発話データを CHAT 形式 (Codes for the Human Analysis of Transcripts system)という共通のフォーマットで電子化し, それらのデータを公開するとともに, コンピュータによる分析ツールである CLAN を提供するというプロジェクトである。Vocd におけるパラメータ D の計算手順は以下のとおりである (McKee, Malvern & Richards, 2000)。

1. テクスト全体からランダムに 35 語抽出。
2. その 35 語での TTR を求める。
3. これを 100 回繰返し，TTR の平均値を求める。
4. 1～3 の操作を N=36,……,50 で行う。
5. 2～4 で求めた 16 の TTR データより, データに最も適合する TTR カーブ推測し，そのカーブに対する D を求める。
6. 1～5 を３回繰り返し，３つの D の平均値を最終的な D のスコアとする。

D は TTR よりも，他の言語発達指標と相関が高く，明らかに熟達度の異なる 2 つのグループをより良く区別することが，いくつかの研究で実証されている。例えば，Chipere, Malvern & Richards (2004) は，英語を母語とする子どもが書いた 918 の物語文を対象に，平均単語長，総語数, スペリングの正しさと D, TTR の相関を調べたところ，D と 3 つの発達指標はすべて有意な正の相関を示したのに対し (それぞれ，r = .59, .45, .29)，TTR と総語数，スペ

リングの正しさは逆に有意な負の相関を示したと報告している (それぞれ r = -.26, -.14)。

Malvern & Richards (2002) は，34 人の英語を母語とするフランス語学習者を対象に，インタビューテストを行い，それぞれの学習者の D や他の発達指標との相関を調べた。結果は，D と異語数・標準化 TTR・流暢さ・インタビューテストのスコアとの相関はすべて有意な正の相関であった (それぞれ，r = .35, .59, .33, .34)。

また Daller & Xue (2007) は，外国語として英語を学習する中国人大学生グループとイギリスの大学で勉強中の中国人留学生グループ計 50 人を対象に，発話データにおける D や TTR などの指標を比較したところ，TTR では 2 群に有意差がなかったのに対し，D では有意差が見られ，イギリスの大学に留学中の中国人の D の方が，外国語として英語を学習する中国人大学生の D よりも有意に高い結果となった。C テストでも同様の結果が得られたことから，D は TTR と比較し，よりよい発達指標と考えられる。

D の信頼性について，McKee, Malvern & Richards (2000) は，CHILDES のデータベースから抽出した 38 人の子どもの発話データを対象に，テクストにおける偶数番目の語のみで算出した D と，奇数番目の語のみで算出した D の相関を調べたところ，r = .76 であったと報告している。2 分した時のテクストは平均 158 語であったことから，D は短いテクストでも信頼性が高いと結論づけている。

D は TTR・テクストの長さをそろえた TTR・標準化 TTR・Guiraud Index・Herdan Index などと比較し，より妥当な語彙の多様性指標だと考えられるが，問題点も指摘されている。Jarvis (2002) は，テクスト中の単語は単なる偶然により並んでいる訳ではないため，Malvern らのランダムに単語を抽出する方法は適切でないと批判している。また McCarthy & Jarvis (2007) は，D はテクスト中のすべての 35～50 語の組み合わせに基づいて算出されるわけではないため，D を算出するたびにスコアがわずかながら異なる点を指摘している。彼らはまた，D がテクストの長さの影響をあまり受けずに，比較的安定したスコアを返すのは，総語数が 100～400 語，200～500 語，250～666 語など，一定の範囲内に限られ，全体的な傾向としては，テクストの長さの増加にともない，D のスコアは少しずつ増加することを実証した。

また，D が明らかに熟達度の異なるグループをうまく区別しない事例も報告されている。Malvern & Richards (2002) は，34 人の英語を母語とするフランス語学習者を対象にインタビューテストを行い，学習者のテクストとフランス語教師であるインタビュアーのテクストにおける総語数と D を調べたところ，総語数はインタビュアーの方が多かったのに対し，D は学習者の方が有意に高かった。また Jarvis (2002) は，英語を外国語として学習するフィンランド人 5・7・9 年生 140 人とスウェーデン人 7・9 年生 70 人，および英語母語話者 5・7・9 年生 66 人を対象に，無声映画を見せ，映画の内容を表す物語文を書かせたところ，フィンランド人 7・9 年生とスウェーデン人 9 年生の学習者グループの方が，英語母語話者 7 年生グループよりも，産出したテクストにおける D が高い結果となった。

2.7 HD-D

McCarthy & Jarvis (2007) は，D の本質はテクストにおける各タイプの出現確率の和（sum of probabilities: SOP）であるとし，D のランダムサンプリングよりも SOP の方が，D が本来測ろうと意図しているもの正確に測ることができると主張した。SOP の算出方法を以下に述べる。例えば，100 語のテクスト中，the が 10 回使用されているとすると，35 語のサンプリング中 the が一度も現れない確率は，0.01034 と算出できる。逆に，35 語中 1 度でも the が現れる確率は，1-0.01034 より 0.98966 となる。このような計算をすべての異なり語に対して求め，足し上げたものを，SOP-35 とする。つまり SOP-35 は，テクスト中のすべての 35 語の組み合わせにおける異語数の平均値と等しくなると考えられる。D は 35-50 語の範囲でサンプリングを行うため，35-50 語の真ん中を取り，SOP-42 に基づく指標が McCarthy & Jarvis(2010)の提案する HD-D である。つまり HD-D は，テクスト中のすべての語の組み合わせに基づいてスコアを算出する D の改良版と言える。

McCarthy & Jarvis (2007, 2010) は，HD-D と D の相関は高く，r = .91～.97 であることを示した。HD-D と D は似たような結果を返すが，完全に相関しないのは D にはランダムサンプリングによるノイズが含まれるためと述べている。彼らはまた，1 つのテクストから取った 100 語について，SOP-42 を求め，101 語目に新出語が表れた場合・既出語が表れた場合で値の変化を調べたところ，既出語が使用された場合，SOP はわずかに減少するのに対し，新出語が使用された場合は比較的大きく増加する。つまり総合的には，テクストの長さの増加にともない，SOP は増加することを示した。したがって，D も HD-D と同様の傾向を持ち，テクストの長さの影響を受けると考えられる。

2.8 Measure of Textual Lexical Diversity (MTLD)

McCarthy & Jarvis (2010) は，D よりも信頼性の高い語彙の多様性指標を目指し，MTLD を開発した。MTLD は，主題に関する飽和の概念を利用する。例えば，話者がある主題について発話するとき，その主題に関する語彙を使い

果たせば異語数が増加することはなく，TTR は減少の一途をたどる。このように異語数が停滞状態に入るポイントを主題に対する飽和状態とし，テクストで使用される語彙が多様であれば，飽和状態に入るまでに長いテクストを要するのに対し，語彙が多様でなければすぐに飽和状態に達する傾向に着目した。MTLD は，一定レベルの語彙の多様性（具体的には TTR = 0.72）を維持するために必要なテクストの平均的な長さから算出される。例えば，*of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) *for* (.714) *the* (.625) *people* (.556) …の場合，people のところで TTR 値が設定の.72 を下回っている。以下のように，people の直後で FACTOR のカウントを 1 増やし，TTR 値をリセットする。

of (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) |||**FACTORS = FACTORS + 1**||| *for* (1.00) *the* (1.00) *people* (1.00)

端数のテクストについて，テクストの最後で TTR が.72 達しない場合のファクター数をどう数えるかという問題がある。TTR = 1 から.72 の差を 1 と考えると，例えば TTR が.887 で終わった場合の差は 0.404 となる。McCarthy らは，この場合のファクター数を 0.404 と考え，それまでのファクター数と合算することにした。テクストにおける総語数をファクター数で割ったものが MTLD のスコアとなるが，予備調査の結果，内的一貫性が低かったと述べている。そこで彼らは，テクストの最初から最後に向かって算出した MTLD のスコアと，テクストの最後から最初に向かって算出した MTLD のスコアの平均値を最終的なスコアとすることとした。

McCarthy & Jarvis (2010) は，6 つの観点から MTLD の妥当性検証を行った。まず，収束的妥当性 (convergent validity) として，MTLD と他の語彙の多様性指標（D, Maas, Yule's K, HD-D）との相関が高いことを示し(r = .69～.85)，弁別的妥当性 (divergent validity) として，MTLD と TTR の相関は中程度（r = .32）であることを示した。また，一貫性の高いテクストは語彙の繰り返しが多く，語彙の多様性は低下することが予想されることから，もう一つの弁別的妥当性として，MTLD とテクストの一貫性指標の間に有意な負の相関（r = -.36）があることを示した。また，内的妥当性 (internal validity) として，MTLD とテクストの長さは相関がないことを報告した（r = -.016）。同じ分析に対し，他の指標ではテクストの長さと有意な相関があったことから (r = .11～.81)，MTLD は既存の語彙の多様性指標よりもテクストの長さから独立であることを示した。彼らは最後に，2 つの観点から増分妥当性 (incremental validity) を検証した。 1 つ目として，レジスターの異なる 16 のテクストにおける MTLD のスコアに有意差があることを示した。2 つ目としては，一貫性の高いテクストは語彙の多様性が低く，逆に一貫性の低いテクストは語彙の多様性が高いと予測し，反復測定分散分析を使って検証した結果，MTLD, TTR, Maas が有意に 2 群を区別したと報告した。

3. 語彙の多様性指標の問題点と今後の課題

語彙の多様性指標はテクストの長さの影響を受けるという信頼性の問題があり，この問題を解決するためのさまざまな試みが続けられてきたが，2 節で概観したように，多くの指標はこの問題が克服できていないのが現状である。この点で最も有望なのは MTLD であると考えられるが，McCarthy & Jarvis (2010) が信頼性の検証のために使用したのは 2000 語に揃えた 144 のテクストであった。第二言語習得研究では，データとなる学習者のテクストは通常 2000 語よりもずっと短いため，より短いテクストにおける指標の信頼性を調査する必要がある。

語彙の多様性指標がテクストの長さに依存する問題を解決したとしても，スコアを語彙発達や一般的な言語発達を表すものとして解釈してよいかという妥当性の問題は残る。McCarthy & Jarvis (2010) の使用したデータは母語話者の書き言葉コーパスであり，第二言語習得研究における語彙や言語発達指標としての MTLD の妥当性は検証されていないと言える。

語彙の多様性指標の中で，比較的テクストの影響が少ないと考えられる D であっても，明らかに熟達度の異なるグループをうまく区別しなかった。この問題について考察を行う。Van Hout & Vermeer (2007) は，32 人のオランダ語を第二言語とする子どもの発話データを分析した Vermeer (1986) の結果より，語彙テストの得点が高かった子どもと低かった子どもの TTR は高い傾向にあり，中間だった子どもの TTR は中間であったことから，TTR 曲線は，発達的に U 字曲線 (U-shaped curve) を描くと述べている。この原因として，冠詞などの機能語の習得が関係していると指摘している。冠詞などの機能語は高頻度で使用されるため，機能語が脱落したテクストにおける TTR は高くなる傾向がある。機能語が産出できるようになるにつれて，TTR は逆に減少する傾向を持つ。機能語がある程度習得され，語彙が豊富になるにつれて，TTR は増加傾向に転じる。D は TTR 曲線全体を表し，テクストの長さの影響を受けにくいという利点があるが，結局は TTR に基づくため，同様の傾向を持つと考えられる。

これまで提案された語彙の多様性指標は総語数と異語

数のみに基づくが，それが妥当なのかという問題がある。異語数を算出する場合には, 形が異なればすべて別の語としてカウントする。例えば，watch, watches, watched, watching, book, books はそれぞれ別の異なり語となる。しかし，watch や book を産出できる学習者は，watching や books を産出するのも容易であると考えられるため，学習者の発表語彙を分析する基準として, 異なり語に基づくのは適切なのか検討する必要がある。

また，Meara & Bell (2001) は，ある写真を見て “The bishop observed the actress.” と描写する学習者と，“The man saw the woman.” と描写する学習者は，語彙の多様性では区別されないが, 前者が入門期の学習者により発せられることはまずないと指摘している。Meara らは，語彙の多様性指標は学習者の作文内で完結する指標 (Intrinsic Measures of Lexical Variety) であるのに対し，外的基準に照らして学習者の語彙を評価する語彙の豊かさ指標 (Extrinsic Measures of Lexical Richness) の必要性を主張している。語彙の多様性指標では，すべての単語の「重み」は等しいと想定されている (Daller, H., & Xue, H, 2007) のに対し，Laufer & Nation (1995), Meara & Bell (2001), Daller, van Hout & Treffers-Daller (2003) らは，語彙の一般的な頻度と習得順位は関係があることから, 単語の頻度を考慮した語彙の豊かさ指標の必要性を説いている。

また，形式のみに基づく語彙の多様性指標では，単語の意味を扱うことはできないため, ある語を適切な意味で使っているかどうかは評価されない。また，高頻度語の多くは多義語であるが, 多義語の中核的な意味より周辺的な意味の方が，習得が難しいと考えられている（Schmitt, 1998; Verspoor & Lowie, 2003）。しかし, 語彙の多様性指標では，学習者が多義語を中核的な意味で使用しても, 周辺的な意味で使用しても区別することができない。

さらに, 語彙の多様性では単語のレベルの評価にとどまっているため, ある語が文法的に適切な環境で使用されているか, 慣用的な連語の観点から適切かなどの問題は扱うことができない。語彙の多様性指標で扱っているのは主に語彙知識の広さと考えることができるが, 今後は語彙知識深さについても考慮していく必要があるだろう。

4. 結論

本研究では，代表的な語彙の多様性指標を取り上げ，それぞれの特徴や問題点について考察を行った。語彙の多様性指標にはテクストの長さの影響を受けるという信頼性の問題があり, 多くの指標はこの問題が克服できていないことが明らかになった。この点で最も有望なのは MTLD であると考えられるが, McCarthy & Jarvis (2010) は母語話者の大規模なデータを使用したため, より短い学習者のテクストにおける信頼性や, スコアを語彙発達や一般的な言語発達を表すものとして解釈してよいかという妥当性の問題を今後検討する必要がある。

語彙の多様性指標がこのような信頼性の問題を克服したとしても残るさらなる課題についても考察を行った。具体的には，機能語習得の影響，語彙の難易度や単語の意味の問題，単語レベルを超えた語彙使用の適切性について，今後どのように考慮し, 評価するのがいいのか検討する必要がある。語彙の多様性指標で扱っているのは主に語彙知識の広さと考えることができるが, 今後は語彙知識深さについても考慮でき, より妥当な発表語彙の発達指標を開発する必要がある。

注

本稿は，2010 年 3 月に名古屋大学大学院国際開発研究科に提出した筆者の博士学位論文『英語学習者の産出語彙における語彙の豊かさ指標 S の提案と論証による S の妥当化』の第２章「総語数と異語数に基づくさまざまな語彙の多様性指標とその問題点」を基に加筆・修正を加えたものである。

引用文献

Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 14-28). Colchester, England: University of Essex.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Chipere, N., Malvern, D., & Richards, B. (2004). Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 137-147). Amsterdam: J. Benjamins.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics, 24* (2), 197-222.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). Cambridge University Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second*

Language Writing, *4*, 139-155.

Guiraud, P. (1954). *Les caractéres statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague, The Netherlands: Mouton & Co.

Hess, G., Haug, H., & Landry, R. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research, 32*, 536-540.

Hess, G., Sefton, K.., & Landry, R. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research*, *29*, 129-134.

石川慎一郎 (2008). 「英語学習者の語彙知識と語彙産出：構造方程式モデリングを利用した英語エッセイコーパスの解析」『統計数理研究所共同研究リポート 215：学習者コーパスの解析に基づく客観的作文評価指標の検討』 15-28.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19* (1), 57-84.

Johnson, W. (1944). Studies in language behavior: 1. A program of research. *Psychological Monographs, 56*, 1-15.

Koizumi, R. (2005). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET Journal*, *9*, 5-33.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced language learner. *Modern Language Journal, 75* (4), 440-448.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19*, 255-271.

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics, 16* (3), 307-322.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learner's written English*. Malmö, Sweden: CWK Gleerup.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon, England: Multilingual Matters.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19* (1), 85-104.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment.* Basingstoke, England: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24* (4), 459-488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42* (2), 381-392.

McKee, G., Malvern, D. D., & Richards, B. J. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*, 323-337.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16* (3), 5-19.

Nation, I. S. P. (2007). Fundamental issues in modeling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp.93-115). Cambridge University Press.

Richards, B. J., & Malvern, D. D. (2000). Accommodation in oral interviews between foreign language learners and teachers who are not native speakers. *Studia Linguistica, 54* (2), 260-271.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*(2), 281-317.

Van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp.93-115). Cambridge University Press.

Vermeer, A. (1986). *Tempo en struktuur van tweede-taalverwerving bij Turkse en Marokkaanse kinderen* (Success and structure in SLA of Turkish and Moroccan children) Unpublished doctoral dissertation). Tilburg University, the Netherlands.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17* (1), 65-83.

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language Learning, 53*(3), 547-586.

（提出日　平成 25 年 1 月 11 日）