LSAを用いた語彙力推定の可能性の探求 ―英語母語話者の語彙サイズと潜在意味解析から予測される語彙サイズとの比較

吉　井　　誠

概要

本研究では、潜在意味解析 (“Latent Semantic Analysis”, 今後本研究では略称 “LSA” を使用）による語彙サイズ推定の可能性について探求する。Coxhead, Nation, & Sim (2015) に記載されている英語母語話者の語彙サイズテスト（Vocabulary Size Test）の得点を参考に、英語母語話者の読書コーパスを用いたLSAから推定される語彙サイズテストの得点と比較した結果から、英語母語話者の得点は推定値よりも高く、LSA分析では実際の受験者の得点を過小評価する傾向があることを示す。一方で、学年が上がることによって語彙サイズが大きくなるパターンには受験者データと推定値に類似性が見られ、LSAを用いて語彙サイズを推定することが有効である可能性が示された。過小評価の現象を考慮しながらLSAを用いることで、読書コーパスから示される読書量を通して、その人の語彙サイズを予測することが可能かもしれないことが分かった。今後、過小評価の原因を追求しながらも、対象者を英語母語話者から第二言語学習者へと広げ、第二言語学習者が読書を通してどのくらい語彙を増やすことができるのか検討する必要がある。

１．はじめに

これまで筆者は潜在意味解析（LSA）を語彙習得研究に用いる可能性について考察してきた（吉井, 2019, 2020）。吉井（2019）ではLSAがTomasello（2003）の説いた用法基盤モデルに根ざしたものであり、言語の用法に頻繁に触れることで言葉を学んでいく理論的背景について説明した。また、具体的なLSA分析の手続き・方法について猪原・楠見（2012）を参照し説明を加えた。そして、これまで行われてきた主要な研究を、とくに語彙習得研究に関連するものを中心に紹介した (Biemiller, Rosenstein, Sparks, Landauer, & Foltz, 2014; Hamada, 2014, 2015, 2017; Landauer & Dumais, 1997; Landauer, Kireyev, &

Panaccione, 2011; 名畑目 , 2012)。そして、LSA 分析の一例として、Vocabulary Levels Test、特に Academic Word Level のテスト問題を取り挙げ、分析を行った。

吉井（2020）では、日本人大学生の 1 年間における語彙知識の変化について、Vocabulary Levels Test を使用して事前事後テストを実施し比較した。また、1 年後の事後テストの結果と LSA から導き出された推定値との比較を行った。LSA はコロラド大学の LSA 分析サイトを利用し、英語話者のコーパスを基に推定した。その結果、2,000 語レベル、AWL (Academic Word レベル) では日本人大学生の得点は母語話者のコーパスから推定される LSA の得点よりも高い結果となった。しかし、3,000 語、5,000 語、10,000 語のレベルになると、英語話者の LSA 推定値の方が日本人大学生の値よりも高くなっていた。全体的には、日本人大学生の得点パターンと英語話者を仮想した LSA の推定値の間には有意な差は見られず、LSA を語彙習得研究に使用可能であることが示唆された。しかし、この調査において、2,000 語レベル、AWL において日本人大学生の方が得点が高いこと、3,000 語、5,000 語レベルにおいて、LSA の得点が高いものの、それほどの差が見られたなかったことから、LSA の推定値は実際の受験者の語彙のレベルを過小評価しているのではないかという疑問が残った。

２．目的と研究課題

本研究の目的としては吉井（2020）で観測された LSA 分析の過小評価の傾向について、Vocabulary Size Test を用いて同じような現象が現れるのか調べることとした。実際の受験者のデータを Coxhead, Nation, & Sim (2015) から用い、LSA の推定値とを比較する。

研究課題としては、英語母語話者の語彙サイズテストの結果と英語話者コーパスに基づく LSA による推定値に違いはあるかを探る。特に以下の 2 点に絞って調査する。1）この二つの全体的なテスト結果に違いはあるか？ 2）英語母語話者の学年間の違いによる語彙サイズの差と LSA 推定得点に表される学年間の差（語彙サイズの差）に違いはあるか？

３．研究材料と方法

3.1　英語母語話者のデータ

英語母語話者のデータは Coxhead et al. (2015) を使用した。この研究の参加者はニュージーランドの８つの学校に在籍する英語母語話者、13 歳から

18歳までの生徒243人で、9年生、10年生、11年生、12年生、13年生という５つの学年の生徒であった。学校で読書活動を行うのに十分な語彙力を備えているか調査することが目的であった。テストは印刷物によるテストとコンピュータ上で受験する方法の二つがあり、学校の環境によって使い分けられた。テスト自体は生徒1人ずつ個別に実施された。結果、語彙サイズテストの得点は　9年生で55.5点 (11,100語)、10年生で57.8点 (11,560語)、11年生で58.1点 (11,620語)、12年生で66.7点 (13,340語)、13年生は66点 (13,200語) という結果であった。

3.2　語彙サイズテスト

語彙知識を測るテストにVocabulary Levels Test (VLT) とVocabulary Size Test (VST) の二つのテストがある。VLTは2,000語、英語圏の大学で遭遇する単語、Academic Word List (AWL)、3,000語、5,000語そして10,000語と5つのレベルに関する受験者の語彙知識を調べている。このテストでは語彙サイズというよりも、各レベルの単語の知識がどれくらいなのかを推定するものである。一方、VSTは語彙サイズを測るものであり、VSTには1,000語レベルから14,000語までをカバーした14,000語テストと20,000語までをカバーしたテストがあり、Coxhead et al. (2015) では後者を使用している。本研究でも同様に後者を採用することとした。

この20,000語テストには1,000語ずつのレベル（1,000語から20,000語まで）が20レベルあり、そこから目標単語を抽出し全部で100問のテストを形成している。1問正解で200語相当の語彙数があると推定しており、100問中の正解数に200を掛けた値が語彙サイズとなる。このテストはL1学習者（英語母語話者）、L2学習者（英語を第二言語として学んでいる者）、両方の学習者の書き言葉の受容的な語彙知識を測る事を目的としており、特に、読解活動などに必要な語彙知識を測るものである。

テスト形式は多肢選択であり、目標とする単語（例：weep）とその単語を用いた短い文章（例：He wept.）が提示される。この文章は単語の品詞などへのヒントにはなるものの、目標単語の意味を推測されないように工夫されている。そして４つの語句（例：① finished his course; ② cried; ③ died; and ④ worried) が選択肢として提示され、受験者は目標単語の意味に最も近いと判断される語句（例：② cried）を選ぶようになっている。

3.3　コロラド大学LSAウェブサイト

本研究のデータ分析を行うためにコロラド大学のLSA研究室で開発されたLSA分析サイト (http://lsa.colorado.edu/) を利用した。このサイトには、

Near Neighbors, Matrix Comparison, Sentence Comparison, One-to-Many Comparison, Pairwise Comparison などの LSA 分析のためのツールが提供されている。本研究では、One-to-Many Comparison というツールを使用し、一つの目標語と複数の単語や語句との、それぞれの関連性について分析する。このサイトの詳しい情報は Dennis (2007) で紹介されている。使用できるコーパスも様々に用意されており、本研究では英語母語話者が出会うと想定される一般的な読み物を集めたコーパス、Touchstone Applied Science Associates Corpus (TASA Corpus) を使用している。この TASA Corpus は 3 年生、6 年生、9 年生、12 年生の 4 つの学年のコーパス、並びに、大学 1 年生相当レベルのコーパスを含んでいる。累計式となっており、例えば、3 年生のコーパスは 3 年生までに英語母語話者の子供たちが目にする教科書や読み物を集めたコーパスであり、6 年生のコーパスは 3 年生のコーパスも含め、6 年生までに生徒が読むであろうと推定されるコーパスを意味する

分析の際は、One-To-Many Comparison のページの Main Text の所に目標単語（例：weep）を入力し、4 つの選択肢の語句（例：① finished his course; ② cried; ③ died; and ④ worried) を Texts to compare の欄に入力する。そして Select a text space において分析の土台となるコーパスを指定し（例：9 年生用のコーパス) 分析する。その結果、それぞれの選択肢に対して目標単語との関連性が数値（コサイン）で現され（例：① finished his course = -0.02 ; ② cried = 0.33; ③ died = 0.31; and ④ worried = 0.10)、数値が一番高いものが LSA で一番関連性が高いと判断された語句である。この例では 4 つの選択肢のうち、「② cried = 0.33」が最も高く、LSA が判断した解答となる。この例の場合、実際のテストの正解も「② cried」であり、LSA が正しい分析をしたことになる。

3.4　母語話者のデータと LSA データ

母語話者のデータは、9 年生、10 年生、11 年生、12 年生、13 年生の５つの学年の生徒からのデータであった。一方、LSA データは先に述べたようにコロラド大学のサイトの TASA Corpus 使用し、そこには 3 年生、6 年生、9 年生、12 年生、大学 1 年生相当レベルのサブコーパスが存在していた。本研究のデータ分析の際は、母語話者のデータと接点のある、9 年生、12 年生、大学 1 年生相当レベルの３つを対象とした。大学 1 年生相当と母語話者のデータの 13 年生と、年齢的に見ても両方とも 18 歳ぐらいを想定しており、同じレベルと判断した。母語話者のデータと LSA データの比較をする際に、結果をグラフ化し、二つのデータの相違を視覚的に判断した。

4．結果

表 1 と図 1 に英語話者のデータと LSA データの比較の結果が示されている。語彙サイズテストには 100 問あり、1 語正解で 200 語相当に値する。100 問中の正解であった項目数が得点として提示されており、それに 200 をかけた語彙サイズが括弧の中に示されている。

「母語話者データ」の箇所でも述べたように、実際の受験者のデータは、9 年生は 11,100 語相当、12 年生または 13 年生 / 大学生では 13,000 語相当の語彙サイズであった。一方、LSA ではそれよりもかなり低い数値が表されている。9 年生は 8,600 語語相当、12 年生では 9,200 語、13 年生 / 大学生では 10,400 語相当の語彙サイズであった。このように全般的に LSA のデータは実際の英語話者のデータを下回っており、吉井（2020）で指摘した LSA の過小評価の傾向を再認する結果であった。

学年間の違いに関しては、母語話者のデータでは、9 年生と 12 年生の間に 11 点（2,200 語ほど）のはっきりとした差が見られたが、12 年生と 13 年生の間には差は見られなかった。一方 LSA データでは、9 年生と 12 年生の間に 3 点（600 語程度）の差が見られ、12 年生から大学 1 年生においては 6 点（1,200 語）の差があった。

表 1　英語話者データと LSA データの比較

	9 年生	12 年生	13 年生 / 大学 1 年生
英語話者	55.5（11,100 語）	66.7（13,340 語）	66.0（13,200 語）
LSA	43（8,600 語）	46（9,200 語）	52（10,400 語）

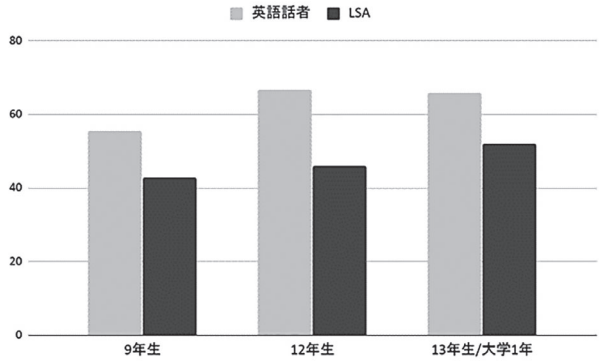

図 1　英語話者データと LSA データの比較

５．考察

分析の結果、母語話者のデータとLSAデータではかなりの差があり、一番顕著な12年生おいては、両者の間に20ポイント（4,000語）ほどの差が生じていた。この結果は吉井（2000）でも現れた現象、LSAが実際の人間による得点を過小評価することを追認する形となった。

吉井（2000）でも取り挙げたが、過小評価の要因としていくつかのことが考えられる。VSTの目標単語の中には、それぞれの学年までのコーパスに載っていない単語があった。一番上の大学1年生相当のコーパスでさえ、100語のうち15語 (rouble, counterclaim, talon, augur, aver, bidet, swingeing, effete, rollick, gobbet, cadenza, spatiotemporal casuist, cyberpunk, pussyfoot) はコーパスに存在せず、分析不可能となった。12年生ではその数は全部で24語、9年生においては26語と100問中4分の1が分析不可能となり、その結果LSAでは得点が得られなかった。それに対して、母語話者の場合は、コーパスでは検出されなくても見聞きしたことが合ったり、選択肢から明らかに異なるものを消去しながら答えを探したりなど、何らかの既知知識や問題解決能力などを総動員して解答していたことが予想される。

さらに、LSAでは4つの選択肢のうち、同じ数値のものが複数存在するケースが見られた。その中に正解が含まれる場合は、LSAでははっきりと正解の判別ができなかったとみなし、得点を与えなかった。このようなケースが9年生で4単語、12年生、大学1年生でそれぞれ1単語あった。その場合は、LSAでは得点なしと判断した。

６．結論

本研究は、英語母語話者の語彙サイズと母語話者の読書コーパスを基にLSA分析して推定した語彙サイズとを比較した。その結果は、吉井（2020）で指摘されたLSAの過小評価（実際の参加者による数値よりも低い現象）を再確認するものとなった。この過小評価の原因について今後も調べる必要がある。そしてこの現象を最大限に抑えながら、あるいはこの差を考慮した上で語彙サイズをより正確に推定することが可能かどうか調べる必要がある。

LSAを用いた語彙力の推定は意義がある。学習者の読書経験を反映させるようなコーパスを構築し、そこからLSAを用いて学習者の語彙力を推定できれば、学習者の持つ問い、どれくらいの量を読んだらどんな力がつくのか、どれくらいの単語を学べるのか、に答えることができる。LSAの活用が語彙習得研究、英語教育へ貢献することが期待される。

謝辞　この研究は JSPS 科研費基盤研究（C）課題番号 18K00748 によって助成を受けている。

参考文献

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T., & Foltz, P. (2014). Models of vocabulary acquisition: Direct tests and text-derived simulations of vocabulary growth. *Scientific Studies of Reading, 18*, 130-154.

Coxhead, A., Nation, P., & Sim, D. (2015). The vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies, 50*(1), 121-135.

Dennis, S. (2007). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (p. 57–70). Lawrence Erlbaum Associates.

Hamada, A. (2014). Using latent semantic analysis to promote the effectiveness of contextualized vocabulary learning. *JACET Journal, 58*, 1-20.

Hamada, A. (2015). Improving incidental L2 vocabulary learning with latent semantic analysis. *ARELE, 26*, 61-75.

Hamada, A. (2017). Estimating input quantity for L2 vocabulary acquisition: A preliminary study of statistical language analysis. *JACET Journal, 61*, 109-129.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*, 92-108.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*, 9-12.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

猪原敬介・楠見孝（2012）「読書習慣が語彙知識に及ぼす影響―潜在意味解析による検討―」*Cognitive Studies, 19*(1), 100-121.

名畑目真吾（2012）「Latent Semantic Analysis (LSA) による空所補充型読解テストの解明―文レベルの意味的関連度を観点として―」*Step Bulletin 24*, 42-58

吉井誠（2019）「潜在意味解析を用いた語彙習得研究の展望について」『熊本県立大学文学部紀要』78 号、45-57

吉井誠（2020）「日本人大学生の語彙知識と英語母語話者の推定語彙知識との比較」『熊本県立大学大学院文学研究科論集』第 13 号、i~xiii.