様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年 １０ 月 ２４ 日現在

機関番号：２７４０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ００７４８

研究課題名（和文）潜在意味解析モデルを用いた語彙学習の予測と妥当性の検証：多量のインプットの役割

研究課題名（英文）Prediction of vocabulary learning through LSA and the study of its validity: Role of extended input

研究代表者
吉井　誠（Yoshii, Makoto）

熊本県立大学・文学部・教授

研究者番号：７０２４０２３１

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：言語習得においてインプットは欠かせない。しかし、どれくらいのインプットを受けるとどのような習得をもたらすのかまだ分からないことが多い。この研究では多量のインプットを主にreadingを中心として受けることにより、語彙がどのように増えていくのか、潜在意味解析を用いて予測する。研究では、実際の学習者のデータとシミュレーションのデータを比較することによってこの予測の妥当性を検証している。

研究成果の学術的意義や社会的意義
潜在意味解析(LSA）を通して、選択肢を基本とした語彙サイズテスト(あるいは語彙レベルテスト）で測定できる語彙の量をある程度の精度で推測できることが分かった。LSAから推定される語彙量と実際の学習者の語彙量を比較したところ、LSAの推定値は学習者の数値を過小評価する傾向にあることがわかった。将来的には、その修正も加えた上での推定がなされるとより正確な推測が可能となると思われる。

研究成果の概要（英文）：We know that a large amount of input is necessary for language learning. Yet we do not know how much amount is needed for what kind of level of learning. This study tackled this question in light of extensive reading and vocabulary learning. This study uses the latent semantic analysis (LSA) to make a prediction of vocabulary learning by analyzing the corpus of input data. The study evaluates the validity of the LSA by comparing the simulated data using the LSA and the actual learner data collected.

研究分野：第二言語語彙習得研究

キーワード：語彙習得　第二言語習得　シミュレーション　潜在意味解析　LSA

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

言語習得においてインプットは必要不可欠である。しかし、どのようなインプットをどれくらい受けるとどのような習得に結びつくのか、研究は進んでいるもののまだ分からないことが多い。そこで、本研究では主に２つの「問い」を設定して研究に取り組むこととした。

一つは、言語習得におけるインプットの役割についてである。本研究では、言語習得を、第二言語語彙習得を中心としながら、多量のインプットを受ける中でどれくらいの量の語彙を学ぶことが可能であるかという「問い」について調べていく。

二つ目の「問い」は言語習得研究におけるシミュレーションの可能性についてである。文系、特に言語習得の領域においてはまだ数が少なく、これからの研究が望まれる。潜在意味解析モデルを用いたシミュレーションを行い、予測の妥当性を検証する。

２．研究の目的

本研究には二つの目的がある。一つは学習者に多読を行う際の指標を示すことである。どの程度の量の教材を読むことにより、どのような学習効果が期待できるのか何らかの指標を提示することを目指す。シミュレーションのデータと学習者データを比較しながら検証する。もう一つは潜在意味解析を用いたシミュレーション研究の妥当性を検証することである。これまで母語話者を対象とした研究が進められてきた中で、本研究は第二言語学習者を対象とする。

３．研究の方法

本研究は当初 3 年計画で目標達成に向けて研究を進めていった。1 年目は準備期間、2 年目は実験期間、3 年目はまとめ並びに総括期間として捉えた。

1 年目は準備期間であり、実験のための多読用教材（Graded Books）の選定を行う。1 年目の後半では、パイロットスタディを行い、少人数の学習者を対象に多読多聴の活動を試行し、研究デザインの調整を行う。同時に潜在意味解析モデルを用いた分析を行い、分析の在り方について検討する。また英語力を測定する方法について検討し、語彙力を図る方法を決める。

2 年目は実験期間であり、授業の一環として多読多聴の活動を実施し、学習者に多量のインプットを与える。潜在意味解析モデルを用いて使用する教材を分析し、語彙学習の予測を行う。予測データと学習者データを比較し予測の妥当性について検証する。

3 年目はまとめの時期であり、2 年目で不足していたデータを追実験により必要に応じて収集する。分析方法などについては、随時、統計専門家の意見を聞く。これまでの研究成果をまとめる。この間、コロナの影響もあり、学会への参加、他の研究者との意見交換が困難な時期が続き、研究期間は当初の 3 年計画から 5 年計画へと変更された。

４．研究成果

（１）潜在意味解析を用いた語彙習得研究の展望について（吉井， 2019）

概要

この論文は潜在意味分析とは何か、なぜ重要なのかについて、先行研究を示しながら説明した。

研究方法

論文の後半では LSA を用いて語彙習得のどのようなシミュレーションが可能であるか、具体例を示しながら解説した。

LSA を用いた研究を具体的に示すために、Vocabulary Levels Test (VLT) を使用し分析した。VLT は学習者の語彙力を 2000 語、3000 語、5000 語、10000 語、Academic Word Level（AWL)と 5 つに分け、各レベルにおいてどれくらいの語彙力があるのか測定している (Nation, 1990: 261-272)。各レベル 10 問からなり、各問題には 6 つの単語と、3 つの意味が表記されている。問題の一つ を例として挙げた。この問題では 6 つの単語 (benefit, labor, percent, principle, source, survey) に対して 3 つの単語の定義 （work, part of 100, general idea used to guide one's actions) が提示されている。学習者はどの単語がどの意味に最も近いか考え選んでいく。

LSA の分析にはコロラド大学が提供しているサイトを用いる(http://lsa.colorado.edu)。

それぞれの定義と単語との類似度を算出し、そこから一番数値が高いものは "principle" (0.16) であるので、これが一番意味的に関連性の高い単語と LSA は分析したことになる。

結果と考察

分析の結果、LSA 分析で正解を正しく推測できたのは 30 単語のうち 20 語（67%）、不正解であったものが 8 語（26%)、2 語(7%）は同じ数値が複数存在したため判断ができなかった (Questions 8 & 10)。今回のシミュレーションによる正答率 67%は Landauer & Dumais (1997) で TOEFL を分析した際の正答率 64.4%を若干上回っており、VLT を用いた分析は LSA の妥当性を示す結果となった。

これからの課題としては LSA に使用する言語コーパスの構築が挙げられる。これまでの研究では英語母語話者を想定したコーパスを利用してきた。今後は、非英語母語話者のコーパスが必要であり、学習者のインプットを反映するようなコーパスを構築していく、コーパス開発が重要となる。

既存の語彙テストを用いシミュレーションを行い、LSA の妥当性の検討を継続していくことも必要である。本論文では VLT の AWL のみで検討したが、他のレベルでも同様の検討が必要となる。Landauer & Dumais (1997) が行ったように TOEFL テストを分析し、日本人学習者に受験してもらい、そのスコアと予測とを比較することも必要である。

（２）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（吉井， 2020）

概要

本研究では、日本人大学生の語彙知識が英文科での 1 年間の学びを通してどのように変化していくのか調べることとした。また、本研究では、大学生の語彙レベルをより客観的に知るために、英語母語話者の大学生の語彙レベルとの比較を行う。ただし、英語母語話者に実際に語彙テストを受けてもらうことは困難であり、ここでは、潜在意味解析という統計手法を用いて母語話者の語彙レベルを推定し比較を行う。

研究方法

参加者は英語英米文学専攻の大学 2 年生であり、1 年後期と 2 年前期に多読多聴を促進するクラスを受講していた。事前・事後テスト、両方を受験した 34 名が本研究の対象者となった。語彙力測定に Vocabulary Levels Test（VLT）を用いた。事前・事後テストの比較は 2000 語、AWL、3000 語、5000 語、そして 1 万語という各レベルで対応のある t *検定*を実施し比較する。英語母語話者と比較するために、最初に LSA を用いて母語話者の大学生の語彙レベルを推定した。本研究では、英語母語話者の大学 1 年生の語彙レベルを測定するために、コロラド大学の LSA サイトで、1 年生のコーパス（"General_Reading_up_to_1st_year_college"）を利用して VLT の各問題を分析した。日本人大学生の事後テストの結果と LSA で推定された英語母語話者の結果の比較にはカイ二乗検定を使用した。

結果と考察

＜日本人学生の１年間の語彙知識の変化＞

図１に各レベルの事前テストと事後テストの結果が視覚的に表されている。全体的には一つのレベル AWL を除き、事前テストより事後テストで得点が伸びていることが分かる。*t* 検定をした結果、統計的に有意な差が検出されたのは、2000 語、5000 語そして 10000 語レベルであった。全体的に語彙のレベルは上がっており、とくに 5000 語や 10000 語など難易度の高い語彙に明らかな変化がみられた。意外な現象としては AWL で伸びが見られなかったことだ。

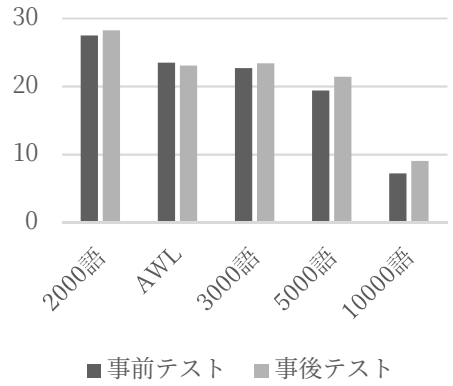

図１　各レベルの事前事後テストの結果

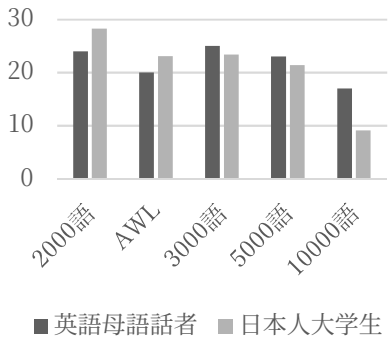

図2　英語話者推定値と日本人実測値

＜LSA で推定される英語母語話者のデータと日本人学生のデータとの比較＞

英語母語話者の推定値と日本人学習者の実測値を比べた結果が図 2 で示されている。予想と反して基本的な語彙のレベル 2000 語と AWL において英語母語話者（Native）の推定値よりも日本人学習者の数値の方が高い結果となった。しかし、語彙レベルが上がるにつれて母語話者の数値が上回っていた。

表 2 の結果をカイ二乗検定で分析したところ、二つのグループの得点のパターンには有意な差は見られなかった（$x^2 = 3.04$, $df = 4$, n.s.）。このことより英語母語話者と日本人学生の語彙知識のパターンには大きな差はみられないとの結果となった。

この研究から大学１年生は日ごろの英文科での学びを通して、AWL を除いて全てのレベルの語彙知識を増やしていたことが分かった。気になったのは、もっと伸びるであろうと期待されていた AWL の数値がそれほど上がっていなかったことである。

英語母語話者の語彙知識を LSA という手法を用いて推定し、日本人大学生の結果とを比較したが、この二つのグループ間には有意な差はないことが分かった。LSA を用いたシミュレーションでは、実際の参加者のデータよりも低い可能性がある。

（３）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，2021）

概要

本研究では、潜在意味解析(LSA）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する。

目的と研究課題

本研究の目的、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について調べることである。実際の受験者のデータを Coxhead, Nation, & Sim (2015) から用い、LSA の推定値と比較する。

研究方法

＜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。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。

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 を使用し、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。

結果と考察

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。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。

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/大学 1 年
Native	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）

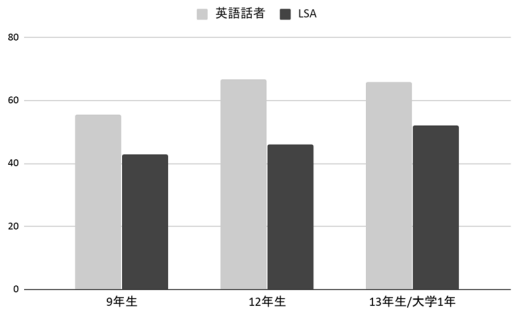

図１　英語話者データと LSA データの比較

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。

成果報告書をしめくくるにあたり

この研究を始めるにあたって主に二つの「問い」を設定して取り組んだ。一つ目は、多量のインプットを受ける中でどれくらいの量の語彙を学ぶことが可能であるかという「問い」であった。実験や観察の結果（吉井，2019；2020)、言語の習得にはインプットが欠かせないことは揺るぎないものであり、インプット受ける中で少しずつではあるが、着実に語彙力が伸びていた。ただし、正確な予測を導き出すまでには至らなかった。二つ目の「問い」は言語習得研究におけるシミュレーションの可能性についてである。LSA を用いて語彙習得を予測することはある程度可能であることが分かった。同時に、LSA の予測は実際の学習者の語彙知識を下回る傾向があることも分かった（吉井，2020；2021)。このような傾向を織り込み、修正した予測値を算出できないか今後も検討が必要である。LSA を用いたシミュレーションなど、これからの言語習得研究においてシミュレーションの役割は益々重要になると思われる。言語習得の現象をいかに具現化してシミュレーションを形成していくかがその大切なステップになると思われる。今回の LSA を中心とした研究もその先駆けとして貢献できることが期待される。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 吉井　誠	４．巻 28
２．論文標題 注の研究の20年の歩み：変化と今後の課題	５．発行年 2022年
３．雑誌名 熊本県立大学大学文学部紀要	６．最初と最後の頁 1~15
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 吉井　誠	４．巻 13
２．論文標題 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	５．発行年 2020年
３．雑誌名 熊本県立大学大学院文学研究科論集	６．最初と最後の頁 i~xiii
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 吉井　誠	４．巻 27
２．論文標題 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	５．発行年 2021年
３．雑誌名 熊本県立大学文学部紀要	６．最初と最後の頁 1~7
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 吉井　誠	４．巻 25
２．論文標題 潜在意味解析を用いた語彙習得研究の展望について	５．発行年 2019年
３．雑誌名 熊本県立大学文学部紀要	６．最初と最後の頁 45-57
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 Makoto Yoshii
2．発表標題 Comparison of vocabulary size test results between learner data and LSA data
3．学会等名 The 31st Conference of JACET Kyushu-Okinawa Chapter
4．発表年 2019年

1．発表者名 吉井　誠
2．発表標題 潜在意味解析を用いた多読教材の分析～語彙習得研究への応用の可能性
3．学会等名 第18回異分野融合テキストマイニング研究会
4．発表年 2019年

1．発表者名 吉井　誠
2．発表標題 多読における読書速度に関する長期的ケーススタディ
3．学会等名 第47回　九州英語教育学会　鹿児島研究大会
4．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関