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　　誠

序論

言語を習得するためにはインプットが必要である。理解できるインプットを多量に受けることが重要である（Krashen, 1985）。しかし言語習得にどれくらいのインプットが必要なのか、多量のインプットを受けるとどのような変化が起こるのか、これらの問いへの答えはまだ見つかっていない。手軽なインプットの手段として近年多読が注目を集めている。読む本さえ手に入れば、これほど手軽に実施できる方法はない。100 万語を読むと英語力が変わると称し関連の書籍が販売されている（酒井 , 2002, 2005; 酒井 , 古川 , 河手 , 2003）。この数値が妥当なものかどうか、信頼のおける目安となるのか、検証は始まったばかりである。これらの問いに応える方法として本研究は潜在意味分析という手法を紹介する。

LSA に関する論文の目的と概要

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ていく。始めに潜在意味分析を支える基本的理論について説明する。先行研究では特に語彙習得研究に関連したものを取り上げ紹介する。潜在意味分析についての理解を促すために論文の後半に具体的な研究事例を示す。最後に今後の研究課題に言及し、語彙習得研究の新しい可能性を示す。 日本語で「潜在意味分析」は「潜在意味解析」とも呼ばれ、英語では “ Latent Semantic Analysis ”、略して“ LSA ”と呼ばれている。本論文でもこれ以降「潜在意味分析（潜在意味解析）」のことを“ LSA ”の略称で表す。

LSA とは何か

LSA とは何かを考える際に、その根本的な理論となっている用法基盤モデル (Usage-based model) について知る必要がある。このモデルでは人が言語習得していくメカニズムとして、言語が実際に使用されている用例にたくさ

ん触れることによって習得していくと説明している。心理学者であるトマセロに代表される理論（Tomasello, 2003）であり、子供が言葉を学ぶ過程を観察することを通して出てきた考え方である。子供は生活の中で親や他者との関わりを通し会話の中で語彙や文法を学習していく。文脈の中で語彙や文法の用法に頻繁に出会うことを通して学んでいく。すなわち、単語の概念を学習することとは、その用法（使われ方）を学ぶことと考える。子供は、幼少時はこのように会話（話し言葉）を通してインプット受け、文字が読めるようになる就学時あたりから本(書き言葉)を読むことを通して多量のインプットを受ける。このようなインプットを受け語彙や文法の用法に触れて知識を増やしていく。

LSA とは、単語や文が他の単語や文とどのように直接的にまたは間接的に登場するか共起表現 (word cooccurence) を統計的に表す方法である。直接的に登場することを直接的共起と呼び、同じ文脈で共起する傾向がある単語同士のことを指し、これらの単語は意味的に類似している。一方、間接的に登場することを間接的共起と呼び、媒介を通して共起関係をつくるものを指している。LSA 以前の分析では直接的共起に焦点を当ててきたが、それを間接的共起まで広げて分析している点が LSA の特徴並びに強みと言える。

LSA の手続き

LSA ではどのような手続きを経て分析を行っていくのであろうか。猪原・楠見（2012: 102-104）は図 1 にあるような概略図を用いて分かりやすく説明している。ここではその要点を記述していく。LSA を構成する主な要素として図 1 の左側に表記されている「出現頻度行列の作成」と右側の「意味空間上での類似度計算」を挙げている。また、出現頻度行列から意味空間上での類似度を計算するために「特異値分解 (singular value decomposition) 」が必要となる。手続きとしては、最初に、対象とする学習者のインプットの現状を代表する言語コーパスが必要になる。このコーパスを用い図 1 の左側の「出現頻度行列の作成」にあるように文脈１、文脈２と一文ずつ並べていく。分析は下線が引いてあるような内容語のみを対象とする（文脈 1 では「魚」「海」「泳ぐ」が内容語）。出現頻度行列とは、単語がどのような文脈で何回出現したかを具体的に行列の中で表現したものである。文脈の中で登場する単語（内容語）を「行」に、テキストにおける意味のまとまり、具体的には文書･記事・段落・文などの文脈を「列」に並べている。出現頻度行列ではそれぞれの文脈において内容語が何回出現したかを記載するが、直接的共起の単語は各文脈内に現れるものがそれに相当する。文脈 1 では「魚」「海」「泳ぐ」、文脈２では「魚」「プランクトン」「食べる」が直接的共起表現となる。LSA

ではこれに加え､「魚」が登場するそれぞれ別の文脈の中で「海」（文脈 1）「プランクトン」（文脈 2）も「魚」という共通の概念を通して間接的につながっていると判断し単語間の関係性の計算に加える。これが間接的共起と呼ばれるものである。

次に「意味空間上での類似度計算」を行っていく。図 1 のように、出現頻度行列は文脈が 20 万列に及び、内容語の行も万を超す単位になることも珍しくない。この膨大なデータから単語間の意味の構造を分析することは困難であるため、効率的に分析する方法として LSA では特異値分解という手法を用いている。これは出現頻度行列を語句ベクトル、特異値、文書ベクトルという３つの行列の積に分解し、不要なものを除いてコンパクトに縮減した行列に変換する方法である（詳しくは豊田（2008: 274-276）を参照）。この方法を次元縮約と呼び、これによって示される行列を意味空間と称する。次元縮約により示される単語間の関係性を類似性 (semantic similarity) と呼び、計算可能な意味空間が作れるようになる。これまでの研究では、約 300 次元が良好な次元縮約と言われている (Landauer & Dumais, 1997)。図 1 の例では出現頻度行列で 20 万に及ぶ文脈があったものを特異値分解で 300 の特徴で縮約している。 概念間の関係性は概念間の距離で計算されるが、距離を算出する方法として概念ベクトル間の角度をコサインで表す。図 1 に示されている「海」「プランクトン」の関係性もコサインで示され、類似度が 2 つのベクトルの角度で示される。コサインの値が 0 に近いほど無関連であり、+1.0 に近いほど関連が強く、概念同士が近い距離にあり意味が類似していると解釈する（コサインの概念､計算方法については豊田（2008：277-279）を、また LSA 全般に関する手続きについての詳細は猪原 (2016: 87-97) を参照)。

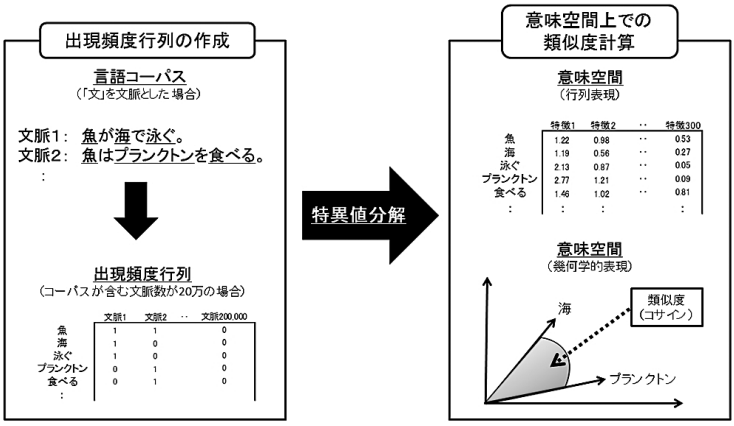

図 1　潜在意味分析の手続きについての概略図（猪原 , 楠見 , 2012: 103)

これまでのLSA研究の紹介

LSAは妥当なのであろうか。これまで先行研究では、LSAを通して示される概念間関連性（類似度）と実際の人間のデータとを比較しながら検証してきた。主に「同義語テスト」「連想課題」「プライミング実験」などを用い、参加者のデータとLSAを通して行ったシミュレーションと比較しながら検証している（詳しくは猪原（2016: 101-145）を参照）。

ここではLSA研究を最初に提唱した先駆け的な研究の一つであるLandauer & Dumais (1997) を紹介する。この研究ではTOEFL (Test of English as a Foreign Language) を使用して研究を行っている。TOEFLはアメリカの大学に留学するために外国人に課せられる英語力を測る試験であり、米国の大学における正規の留学のためにはこのテストの受験が必要である。TOEFLの中の語彙力テストをLSAで分析している。同義語テストの形式をとっており、目標語1語に対して候補語4語が与えられ、受験者は目標語に最も意味が近いと思われる単語を一つ選ぶ。分析のための言語コーパスとしてGrolier's Academic American Encyclopediaという百科事典の電子版を用いている。その中から3万を超す記事、異なり語数として6万語に及ぶ単語を使用している。同義語テストは80問からなり、LSAによる正答は51.5問であり正答率は64.4%であった。実際に留学希望の学生に受験してもらった結果は正答数51.6問、正答率は64.5%と非常に近いものとなった。これはLSAが人間の語彙学習を適切に予想しうる妥当な方法であることを示している。

Landauer & Dumaisは20年ほど前に行われた研究であるが、それ以来LSAを用いた様々な研究がなされてきた。その中でも特に語彙発達に関連する研究について紹介する。LSAを通して、語彙発達をシミュレーションして予測する研究（Landauer, Kireyev & Panaccione, 2011)、実際の参加者のデータと比較しながら予測の妥当性を調べている研究 (Biemiller, Rosenstein, Sparks, Landauer & Foltz, 2014) がある。

Landauer, Kireyev & Panaccione (2011) では語彙の発達を語彙の成熟（Word Maturity）という概念を用い、シミュレーションを通して観察することを試みている。英語母語話者の子供達が遭遇すると想定される言語コーパスからLSAを行い、子供の成長に伴い語彙がどのように発達していくのかシミュレーションを通して観察している。語彙発達を観察するには、これまでの研究では主に実際の子供のデータを収集し利用していたが、シミュレーションという新しい手法の登場で研究の可能性が広がった。

Biemiller, Rosenstein, Sparks, Landauer & Foltz (2014) の研究では、どのように子供たちの語彙が発達していくかLSAを用いてシミュレーションし、さらに、実際に子供たちに実施した語彙テストの結果と比較している。その

結果は相関係数で .67 から .74 という高い数値を得ることができ、LSA が妥当であることを証明している。語彙発達のシミュレーション研究を進めていく上で重要な研究である。

次に、最近日本で行われている、英語教育における LSA 研究について紹介する（名畑目 , 2012; Hamada，2014；2015；2017）。名畑目（2012）では英検の空所補充型読解テストについて LSA を用いて分析している。このテストでは空欄のある文を読み、その空欄に何が入るかを受験者は補充しなければならない。研究では文レベルの意味的関連度に着目している。調査の結果、空所に入る単語とその文中の他の単語との意味的関連性が重要であるのみならず、空所を含む文とその前後の文の意味的関連や、段落の中における意味的関連も重要になることが示唆された。文章理解、文の中における単語理解を考える際に、単語間の意味的関連性のみならず、前後の文章や段落の関連性も重要であることを示している。

次に Hamada によって実施された一連の語彙習得に関連する LSA 研究について言及する（Hamada，2014；2015；2017）。Hamada (2014) では 105 名の日本人大学生が意味と例文が共に提示された 20 単語を学習している。例文の中に提示されている目標単語と文中の他の単語との関連性について LSA を用いて分析し、文脈の質が学習に与える影響について調査している。使用したコーパスはコロラド大学の LSA サイトで提供されている General_Reading_up_to_1st_year_college という英語母語話者の大学生レベルのコーパスである。学習方略についても調査を実施している。実験の結果、文中の単語の関連性が高いほど目標単語の学習が促進されることが示された。例文を示しながら単語を学ばせることは重要であるが、その例文の質にも気を配る必要があることを示している。

Hamada (2015) では 153 名の大学生を参加者として 20 単語を学習させている。Hamada（2014）と同じように単語の意味を例文と共に与えている。この実験では、目標単語と例文のほかの単語との意味的関連性の高いものと低いものを LSA を用いて区別し、付随的語彙学習において、どちらのグループの単語がよりよく習得されるかについて調査している。その結果、LSA により関連性が高いと判断された文脈の単語の方がよりよく学習されていたという結果を残している。Hamada (2014) で示された例文の質の重要性について実験を通して再認する結果となった。

Hamada（2017）では英検の問題を用い、LSA を通して解答しているシミュレーション研究である。この研究でもコロラド大学が提供する LSA サイトを用い、そのサイトで提供されている 異なる学年のコーパス（3 年生、6 年生、9 年生、12 年生）、そして最後に General_Reading_up_to_1st_year_college

という大学生対象のコーパスを用いて分析している。対象とした英検は4級から準1級まで5種類であり、それぞれのレベルの語彙問題に対してLSAを使用して解答した。その結果、学年が上がりコーパスのサイズが増えると英検問題での正解率が上がる様子が観察された。Landauer et al. (2011) やBiemiller et al. (2014) では英語母語話者を対象としてLSAのシミュレーションを行っていたが、Hamada (2017) では対象を広げ日本人学習者を想定してシミュレーションを実施し語彙発達の過程を観察している。

LSAを利用したVLTの分析

LSAを用いた研究を具体的に示すために、Vocabulary Levels Test (VLT) を使用し分析していく。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を測定する (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。表1に問題の一つ (Question 1) を例として挙げLSAによる類似度を掲載している。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。各問題で3つの単語の意味を答え、各レベルで30単語の意味で構成されている。本論文では、5つのレベルの中でも特にAWLを使用する。これは英語母語話者の大学生が大学での学習において頻繁に遭遇する単語を集めたものである。LSAの分析にはコロラド大学のLSA研究室が提供しているLSAサイトを用いる (http://lsa.colorado.edu)。先行研究でも多数このサイトを利用して分析を行っており本論文でもそれを踏襲している。なおこのサイトを利用してLSAを実施する際の手続きについては詳細がDennis (2007）に示してある。

このサイトで、" One-to-Many Comparisons " という分析を選び、必要な設定を行う。参考にするコーパスはHamada（2014；2015；2017）でも使用されていた " General_Reading_up_to_1st_year_college " という、英語母語話者の大学1年生レベルのコーパスである。300次元（Factors）を基準として分析を行った。表1ではAWLのQuestion 1の結果を提示している。

表 1 LSA を用いた AWL の問題（Question 1）の分析

Question 1	work	part of 100	general idea used to guide one's actions
benefit	**0.26***	0.03	0.02
labor	0.20	-0.03	0.00
percent	0.11	**0.13**	0.05
principle	0.13	-0.10	**0.16**
source	0.15	0.05	0.08
survey	0.22	0.04	0.05

表 1 には行に 6 つの単語が並んでおり、列に英語による単語の定義づけが 3 つ掲載されている。そして LSA によって分析された類似値が示されている。定義づけの一つ（“general idea used to guide one's actions”) を例に見ていく。6 つの単語との類似値の中で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。このように LSA が選出した単語は太字で示している。実際に正解も “principle” であり、正解の単語には下線が引いてある。よってこの定義については、LSA は正しく正解を導きだすことができたので、そのような場合には数値は太字と下線が記してある。真ん中の行の定義 (“part of 100”) においても一番高い数値は “percent” の 0.13 であり、これが LSA で分析の結果である。正解でもあるので、表に示されているように太字並びに下線が引いてある。しかし、“work” の意味に関しては、LSA の推測では “benefit” が 0.26 で一番高い数値であるが、正解は “labor” であり、正確に予測することは出来なかった (不正解の場合は数値に * を記している)。この表では Question 1 の結果のみ表示しているが、残りの問題 (Questions 2-10) の結果は APPENDIX に掲載している。10 問の各問題に 3 つの意味の選択肢が用意されているので、AWL 全体としては合計 30 の単語の意味選択を分析したことになる。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 LSA 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 を用いた分析でも LSA の妥当性を示す結果となった。

これからの研究の課題・展望

これまでの先行研究を通して、LSA が妥当なものであることが示されてきた。また、LSA を通して言語発達のシミュレーション研究も活発に行われてきていることなどが分かった。それでは、今後の LSA 研究の課題にはどのようなものがあるだろうか。LSA を通して可能な第二言語語彙習得研究についても言及していく。

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、Grolier's Academic American Encyclopedia に代表されるように英語母語話者を想定したコーパスを利用してきた。コロラド大学の LSA のサイトで使用されているのもこのコーパスである。これは英語母語話者が大学生までに遭遇するであろうと仮定される書籍や論文などのコーパスであり、非英語母語話者が遭遇するコーパスとは異なる可能性が高い。名畑目 (2012:55) も指摘しているように、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どのように構築していくのか、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 LSA で分析し、日本人学習者に受験してもらい、そのスコアと LSA の予測とを比較することも必要である。

最後に、今後の語彙習得研究において LSA をどのように使用していけるのか、その可能性について言及する。非英語母語話者のコーパスが必要であることは述べたが、その開発の一つとして、多読教材をもとに言語コーパスを構築することが考えられる。現在、カリキュラムの一環として多読多聴の授業を行っているが、多読教材のコーパスを構築して LSA を実施することが可能である。例えば、Cambridge English Readers のシリーズの Level 1 の本 10 冊を 1 年次に読ませ、Level 2 の本各 10 冊を 2 年次に読ませる。VLT で事前事後に語彙力を測定する。これらの本を基に作成したコーパスを用い LSA を実施し、VLT の問題を解いてみる。学習者のデータと LSA によるシミュレーションデータを比較し妥当性を検証していく。将来的にはコーパスを広げ、Level 3 から Level 6 までをカバーし、多読教材を読破することによってどれくらい語彙力を伸ばすことが可能か検証したいと考えている。もちろん、そのためにはコーパスを構築する作業のみならず、コロラド大学の LSA サイトで実施しているような分析を独自に実施しなければならない。このような分析を可能とするプログラムの開発も必要となる。R などの言語を用いてそのようなツールを開発することが可能か検討していく。

LSA 研究はシミュレーションによる調査という新たな手法を提供している。語彙習得研究に応用することで新たな発見につながることが期待される。今後の研究の成果が待たれる。

参考文献リスト

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of vocabulary acquisition: Direct tests and text-derived simulations of vocabulary growth, *Scientific Studies of Reading*, *18*(2), 130-154.

Dennis, S. (2007). How to use the LSA website. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.). *Handbook of latent semantic analysis* (pp. 57-70). Mahwah, NJ: 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 L2 vocabulary learning with latent semantic analysis. *JACET Journal, 59*, 61-76.

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

Krashen, S. (1985). *The input hypothesis: Issues and implications*. London: Longman.

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*(1), 92-108

Nation, I. S. P. (1990). *Teaching & learning vocabulary*. Boston, MA: Heinle & Heinle Publishers.

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

Cambridge: Harvard University Press.

猪原敬介（2016）.『読書と言語能力：言葉の「用法」がもたらす学習効果』京都大学学術出版会

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.

酒井邦秀（2002）.『快読 100 万語！ペーパーバックへの道』筑摩書房

酒井邦秀（2005）.『教室で読む英語 100 万語』大修館書店

酒井邦秀、古川昭夫、河手真理子（2003）.『今日から読みます英語 100 万語！』日本実業出版社

豊田秀樹編（2008）.『データマイニング入門』東京図書

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

Appendix: AWL 問題 (Question 2-10) の LSA の結果
*LSA で不正解であったもの、又は判断が困難なもの

Question 2	money for a special purpose	skilled way of doing something	study of the meaning of life
element	0.01	0	0.01
fund	**0.19**	0	0
layer	0.02	-0.02	0
philosophy	0.05	0.05	**0.15**
proportion	-0.06	0.03	-0.02
technique	0.02	**0.08**	0.08

Question 3	total	agreement or permission	trying to find information about something
consent	0.05	**0.39**	-0.06
enforcement	0.04	0.04	0
investigation	0.11	-0.06	**0.11**
parameter	0.19	0	-0.08
sum	**0.55**	0.07	-0.11
trend	0.14	-0.02	0

Question 4	ten years	subject of a discussion	money paid for services
decade	**0.06**	-0.04	-0.16
fee	0.02	-0.06	**0.26**
file	-0.03	**0.11***	-0.02
incidence	0.02	0.02	-0.03
perspective	-0.06	0.05	0.01
topic	0.03	0.07	-0.01

Question 5	action against the law	wearing away gradually	shape or size of something
colleague	-0.01	0.01	0.04
erosion	0.01	**0.31**	-0.03
format	-0.02	0.03	-0.11
inclination	0.03	0.01	-0.06
panel	-0.09	-0.08	**0.05***
violation	**0.31**	0	0.02

Question 6	change	connect together	finish successfully
achieve	**0.24***	0.08	0.02
conceive	0.09	-0.11	**0.03***
grant	0.1	0	0.02
link	0.06	**0.23**	0.06
modify	0.22	-0.07	0
offset	0.04	-0.11	-0.08

Question 7	keep out	stay alive	change from one thing into another
convert	-0.06	0	0.02
design	0.02	-0.01	0.02
exclude	0	-0.01	**0.06***
facilitate	-0.02	0.01	0
indicate	-0.08	-0.04	0.03
survive	**0.08***	**0.16**	0.02

Question 8	control something skillfully	expect something will happen	produce books and newspapers
anticipate	-0.05	**0.08**	-0.09
compile	**0.03***	0.06	0.06
convince	-0.03	0	0.04
denote	0.02	-0.03	-0.13
manipulate	**0.03***	0	0
publish	0.01	-0.07	**0.43**

* 同じ数値が複数あり判断が困難

Question 9	most important	concerning sight	concerning money
equivalent	-0.09	0.08	0.01
financial	-0.05	-0.01	**0.05**
forthcoming	**0.08***	0.15	-0.02
primary	0.05	0.03	0.01
random	-0.04	0.03	-0.03
visual	0.01	**0.18**	0.04

Question 10	last or most important	something different that can be chosen	concerning people from a certain nation
alternative	-0.03	**0.15**	-0.13
ambiguous	-0.08	-0.08	0.04
empirical	-0.05	0.04	-0.11
ethnic	0.02	0.05	**0.08**
mutual	**0.03***	-0.07	-0.06
ultimate	**0.03***	0.05	-0.08

* 同じ数値が複数あり判断が困難