データマイニングによる誤り分析と辞書開発

小林 雄一郎
日本学術振興会

本研究の目的は，アソシエーション分析とクラスター分析を用いて，学習者による誤りの共起関係を明らかにすることである。具体的には，「ある誤りを犯す学習者は，他にどのような誤りを一緒に犯す可能性があるのか」という情報を大量に蓄積し，それらをいくつかの典型的なタイプへと統計的に分類する。

Data mining, error analysis and dictionary

Yuichiro Kobayashi
Japan Society for the Promotion of Science

The present study applies data mining in an effort to examine the co-occurrence patterns of EFL (English as a foreign language) learners' errors. Association analysis and cluster analysis are employed to spotlight frequent patterns of errors.

1. はじめに

電子化コーパスを最も本格的に使って編集された最初の辞書は，1987 年の *Collins COBUILD English Language Dictionary* である。それ以降，辞書編集におけるコーパスの重要性が広く認識されるようになり，現在，英語圏の主要な出版社の辞書は，コーパスに基づくものである (e.g. Rundell, 1998)。

そして，近年，外国語学習者が産出した言語のコーパス (学習者コーパス) から得られた誤り情報を活用する試みが世界的に見られる。たとえば，*Longman Essential Activator* では，Longman Learners' Corpus (LLC) から得られた誤り情報を用いて，(1) コロケーション誤り，(2) 文法誤り・統語誤り，(3) スペリング誤り，(4) 語順誤り，という 4 つの領域に関して，"help box"を設けている。この"help box"という発想は，典型的な学習者の誤りをはっきりと提示し，それが正用ではなく誤りであることを明確にして，誤りを生むことになった内容の正しい表現方法を学習者に示すことにある (Gillard & Gadsby, 1998)。現在，同様のコラムは，*Cambridge Advanced Learner's Dictionary* における"common mistake"，*Macmillan English Dictionary for Advanced Learners* における"Get it right"など，いくつかの学習者向け辞書にも見られる。我が国の例では，『ロングマン英和辞典』が警告の三角印をつけた「エラー・ノート」を設けている。また，*Longman Dictionary of Common Errors* のような，コラムという形式ではなく，学習者の誤りだけを記述した辞書も出版されている。

しかしながら，現在の辞書において，学習者の誤りに関する記述は，非常に限定されたものである。特に，投野 (2012) が指摘しているように，(1) 誤りの頻度情報が考慮に入れられていないために，個々の誤りの重要度が不明であること，(2) 習熟度別のコーパスを使用していないために，学習プロセスを勘案した記述が乏しいこと，などの問題点がある。このような状況において，学習者コーパスにもとづく誤りの統計的分析を行うことは，非常に有意義なことである。

本研究の目的は，学習者による誤りの共起関係を明らかにすることである。具体的には，「ある誤りを犯す学習者は，他にどのような誤りを一緒に犯す可能性があるのか」という情報を大量に蓄積し，それらをいくつかの典型的なタイプへと統計的に分類する。

2. 実験データ

本研究で用いる実験データは，日本人英語学習者の話し言葉コーパスである NICT-JLE Corpus (和泉ほか, 2004) である。

このコーパスには，Standard Speaking Test (SST) にもとづく 9 段階の習熟度情報，そして，47 種類にわたる詳細な誤り情報が付与されている。[1] 表 1 は，実験データの概要である。

表 1　実験データ（人数）

Lv. 1	Lv. 2	Lv. 3	Lv. 4	Lv. 5
1	7	28	43	30
Lv. 6	Lv. 7	Lv. 8	Lv. 9	
28	16	9	5	

また，以下は，コーパスに付与された誤り情報の例である（ボールドになっている箇所が誤りタグ）。なお，これらの情報は，英語に精通した複数の日本人によって付与されている。

<B>Yes. <F>Uh</F>. Usually, **<at odr="1" crr="the"></at>** museum **<v_agr odr="2" crr="opens">**open**</v_agr>** on **<n_num odr="3" crr="Saturdays">**Saturday**</n_num>**and **<n_num odr="4" crr="Sundays">** Sunday**</n_num>**. <SC>And</SC> so **<av_pst odr="3" crr="usually the holiday is on Mondays">**the holiday is **<prp_lxc1 odr="1" crr="on"></prp_lxc1> <n_num odr="2" crr="Mondays">**Monday **</n_num>** usually **</av_pst>**, so I'm off today.</B>

上記の例を見ると，最初に冠詞（at）の脱落型誤りがあって，次に *open* という動詞と主語の一致（v_agr）に関する誤りがあることが分かる。なお，<B>はそれが学習者の発話であることを示し，<SC>と<F>はそれぞれ自己訂正とフィラーを表している。

[1] なお，本研究はパイロット・スタディであり，実験データの規模も小さいことから，習熟度情報は分析に用いていない。

3. 実験手法

3.1. アソシエーション分析

本研究で主に用いる実験手法は，アソシエーション分析である。これは，大量のデータから「X ならば，Y である」という因果関係に関する情報（アソシエーション・ルール）を見つけるための手法で，マーケティングの分野などで，POS システムの購買履歴から「パンとバターを購入した人の 90%が牛乳も購入している」といった情報を抽出するのに用いられている。また，教育分野では，e ラーニングの学習履歴データの分析などに応用されている (e.g. Pahl & Donnellan, 2002; Wang & Shao, 2004)。

図 1　アソシエーション・ルール

アソシエーション・ルールを抽出する際，何らかの評価指標が必要となる。一般に用いられている指標としては，support 値，confidence 値，lift 値がある。

- support (X => Y)＝条件 X と結論 Y を含むデータ数／全データ数
- confidence (X => Y) ＝条件 X と結論 Y を含むデータ数／条件 X を含むデータ数
- lift (X => Y)＝confidence／結論 Y を含むデータ数

本研究では，47 種類の誤りのアソシエーション分析を行い，学習者による誤りの共起情報を大量に集積する。なお，アソシエーション分析の場合，通常の共起分析で用いられる対数尤度比や相互情報量といった指標と異なり，3 つ以上の事象の共起を容易に扱うことが可能である。また，個々の共起項目に関して，単なる相関関係ではなく，どちらが原因でどちらが結果であるかという情報を得られる。

3.2. クラスター分析

クラスター分析とは，個体間の類似度，あるいは非類似度（距離）に基づいて，最も似ている個体から順番に結合して，クラスターを作って行く手法である。クラスター分析には，階層型クラスター分析と非階層型クラスター分析があるが，本研究では前者を用いる。

階層型クラスター分析とは，クラスタリングの結果をデンドログラムと呼ばれる樹形図で表わす。樹形図では，いくつかの個体が階層的に結合されてクラスターを形成し，最終的には複数のクラスターが結合されて 1 つのクラスター（木）となる様子を見ることができる。なお，クラスター分析は，学習者コーパスの分析などに応用されている (e.g. Abe, 2012)。本研究では，アソシエーション分析で抽出されたルールのクラスター分析を行う。

4. 結果と考察

図２は，それぞれの誤りの 10000 語あたりの相対頻度を視覚化したものである。これを見ると，冠詞 (at)，前置詞の語彙選択 (prp_lxc)，動詞の時制 (v_tns)，名詞の単複 (n_num)，動詞の語彙選択 (v_lxc) などが高頻度な誤りであることが分かる。

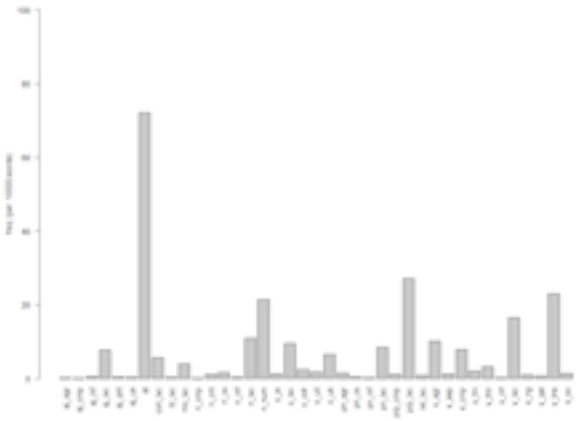

図２　誤りの相対頻度（10000 語あたり）

そして，アプリオリ・アルゴリズム (Agrawal & Srikant, 1994) を用いて，「ある誤りを犯す学習者は，他にどのような誤りを一緒に犯しているのか」というアソシエーション・ルールを抽出した。アプリオリ・アルゴリズムとは，評価指標の下限を設定し，それを下回る組み合わせに関する計算を省略することで，ルール抽出を高速化するものである。本研究では，support 値（全事象中でルールが起きる確率）の下限を 0.1，confidence 値（条件 X が起こったときに結論が Y になる確率）の下限を 0.8，ルールの条件部と結論部を合わせた最大の長さを 5 と設定した結果，142504 個のルールが抽出された。

図 3〜6 は，抽出されたルールにおける support 値，confidence 値，lift 値それぞれの分布，そして，それらの値の関係をまとめたものである。

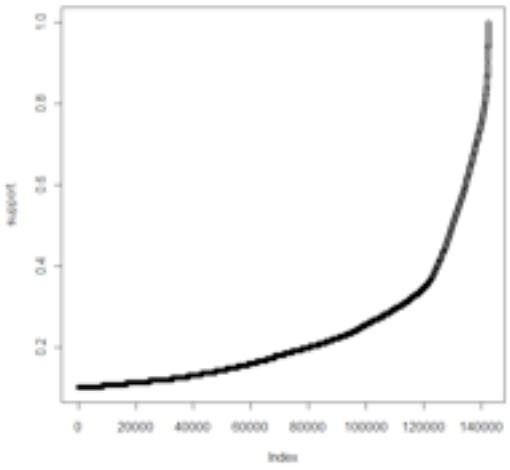

図３　support 値の分布

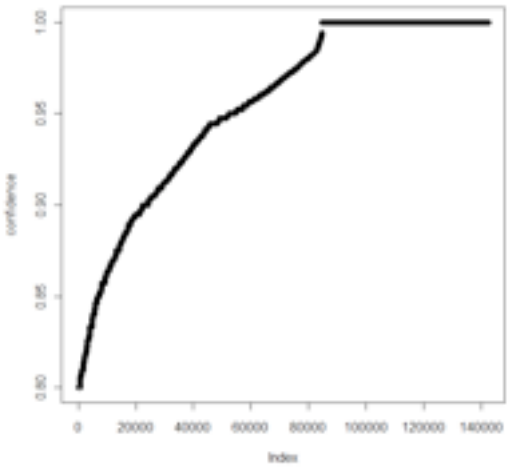

図４　confidence 値の分布

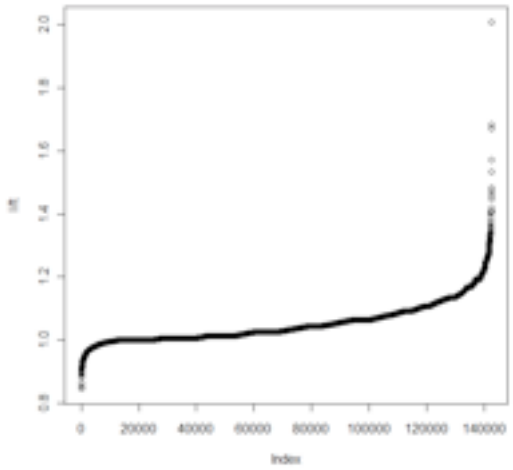

図５　lift 値の分布

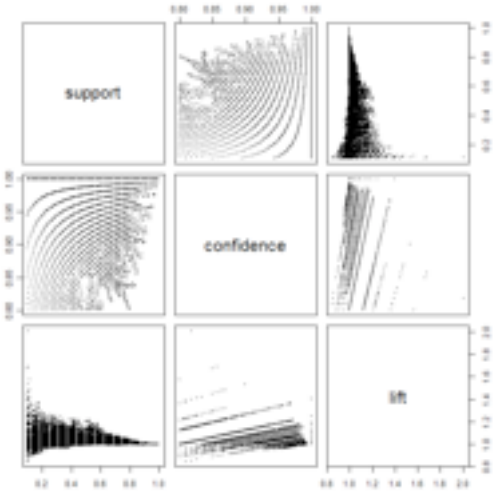

図６　support 値，confidence 値，lift 値の関係

表２は，support 値上位 10 個のルールをまとめたものである。

表２　support 値上位 10 個のルール

rule	support
{} => {at}	1.00
{} => {prp_lxc}	0.99
{prp_lxc} => {at}	0.99
{at} => {prp_lxc}	0.99
{} => {n_num}	0.99
{n_num} => {at}	0.99
{at} => {n_num}	0.99
{n_num} => {prp_lxc}	0.98
{prp_lxc} => {n_num}	0.98
{n_num, prp_lxc} => {at}	0.98

表２によると，support 値が高い誤りは冠詞 (at)，前置詞の語彙選択 (prp_lxc)，名詞の単複 (n_num) で，support 値の最も高いルールは前置詞の語彙選択 => 冠詞 (prp_lxc => at) である。このようなルールは，それ自体で言語学的にも言語教育的にも非常に示唆的なものである。しかしながら，142504 個におよぶルール全てを人間が把握することは困難であり，全てを辞書に記述することは不可能である。そこで，実際の言語教育や辞書編纂に向けて，典型的な誤りの共起パターンのみを抽出し，それらを類型化する技術が必要となる。以下では，その一例として，前置詞の語彙選択誤りに焦点を当てる。

図７は，結果部が前置詞の語彙選択誤りとなるルールのうち，support 値が 0.9 以上のルール 29 個を対象に，クラスター分析（ウォード法）を行った結果である。図中の破線の位置をカッティング・ポイントとした場合，29 個のルールは３ つのクラスターに分けられる。

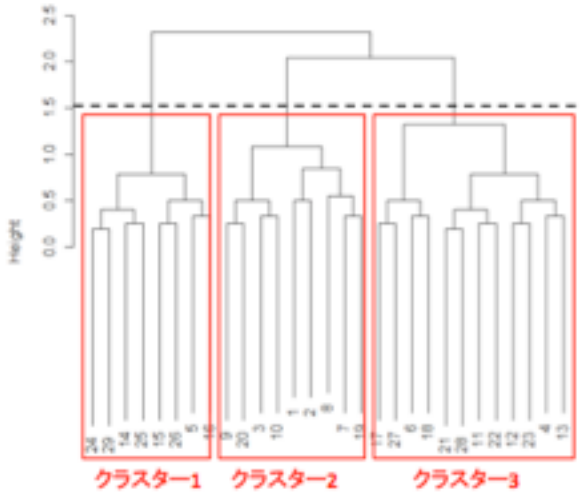

図７　前置詞の語彙選択誤りのクラスター分析

そして，表3～5 は，それぞれのクラスターに含まれているルールをまとめたものである。

表3　クラスター1 に含まれるルール

rule	support
{n_num, v_lxc, v_tns} => {prp_lxc}	0.91
{at, n_num, v_lxc, v_tns} => {prp_lxc}	0.91
{v_lxc, v_tns} => {prp_lxc}	0.93
{at, v_lxc, v_tns} => {prp_lxc}	0.93
{n_num, v_lxc} => {prp_lxc}	0.94
{at, n_num, v_lxc} => {prp_lxc}	0.94
{v_lxc} => {prp_lxc}	0.95
{at, v_lxc} => {prp_lxc}	0.95

表4　クラスター2 に含まれるルール

rule	support
{n_lxc, n_num} => {prp_lxc}	0.91
{at, n_lxc, n_num} => {prp_lxc}	0.91
{n_lxc} => {prp_lxc}	0.92
{at, n_lxc} => {prp_lxc}	0.92
{n_num} => {prp_lxc}	0.98
{at, n_num} => {prp_lxc}	0.98
{} => {prp_lxc}	0.99
{at} => {prp_lxc}	0.99

表5　クラスター3 に含まれるルール

rule	support
{n_num, v_tns} => {prp_lxc}	0.96
{at, n_num, v_tns} => {prp_lxc}	0.96
{v_tns} => {prp_lxc}	0.97
{at, v_tns} => {prp_lxc}	0.97
{n_num, v_agr, v_tns} => {prp_lxc}	0.91
{at, n_num, v_agr, v_tns} => {prp_lxc}	0.91
{v_agr, v_tns} => {prp_lxc}	0.92
{at, v_agr, v_tns} => {prp_lxc}	0.92
{n_num, v_agr} => {prp_lxc}	0.93
{at, n_num, v_agr} => {prp_lxc}	0.93
{v_agr} => {prp_lxc}	0.94
{at, v_agr} => {prp_lxc}	0.94

まず，クラスター1 では，動詞の語彙選択誤り（v_lxc）や動詞の時制の誤り（v_tns）が前置詞の語彙選択誤りと共起している。つまりは，動詞の誤りが引き金となって，前置詞の誤りを引き起こしているパターンである。具体的には，*look at* であるべき箇所を *watch* としている例，*looks like* であるべき箇所を *seems* としている例，*go out* であるべき箇所を *exit* としている例のように，本来は動詞＋前置詞の表現で表すべき箇所を動詞 1 語のみで表現しようとしている例が多かった。一見似たような意味を持つ表現のどちらが適切であるかは，それらの表現が実際に使われる文脈にも依存するため，学習者には判断が難しいことも多い。

次に，クラスター2 では，名詞の語彙選択誤り（n_lxc），名詞の単複の誤り（n_num），冠詞誤り（at）などが前置詞の語彙選択誤りと共起している。つまりは，名詞句の誤りが引き金となって，前置詞の誤りを引き起こしているパターンである。実際の例を見ると，本来は前置詞が必要な箇所に前置詞がない脱落型の前置詞誤りが非常に多く，その中には，*for about two months* であるべき箇所を *about two month* とするような名詞の単複誤りと複合した誤り，*went to a restaurant* であるべき箇所を *went restaurant* とするような冠詞誤りと複合した誤りなどがあった。冠詞も前置詞も日本語には存在しない品詞であるため，話し言葉のようなリアルタイムのコミュニケーションの中では，高い頻度で誤りを犯しやすい傾向が見られる。

最後に，クラスター3 では，動詞の時制の誤り（n_lxc）に加えて，主語・動詞の人称・数の一致に関する誤り（v_agr）などが前置詞の語彙選択誤りと共起しており，やや解釈が難しい。これらは，必ずしも文法的な依存関係のある共起パターンではなく，ある習熟段階にいる学習者が発話内で犯しやすい誤り同士を結びつけているパターンであると思われる。

5. おわりに

本研究では，アソシエーション分析とクラスター分析を用いて，日本人英語学習者の話し言葉における誤りの共起分析を行った。また，結論部が前置詞の語彙選択誤りとなるルールのクラスター分析を行った。その結果，3 つのクラスターが抽出され，クラスター1 では動詞関連の誤り，クラスター2 では名詞句関連の誤り，クラスター3 では動詞の時制や人称・数の一致に関する誤りがそれぞれ顕著であった。

今後の課題は，主に 3 つある。第 1 の課題は，共起のスパンの問題である。今回の実験で，やや解釈が難しいクラスター3 が抽出されたのは，同一話者が同一のインタビュー内で犯した誤りを共起と見なしていることが一因である。共起を抽出するスパンをインタビューよりも短い単位（e.g. ターン，文）とした場合，文法的，あるいは意味的な結びつきの強い共起を抽出しやすくなるはずである。

第 2 の課題は，誤りタグ付与のガイドラインの見直しで

ある。今回は，実験データである NICT-JLE Corpus の誤りタグをそのまま用いたが，教育的な応用という観点では，より詳細な誤り情報の付与が望ましい。また，誤り情報の付与にあたっては，自然言語処理における誤りの自動検出技術 (e.g. Izumi, *et al.,* 2004; Nagata & Nakatani, 2010) を部分的に用いることができる。

第 3 の課題は，習熟度別の分析である。そのような分析を行うことによって，初級者に特徴的な誤りのパターンは何か，上級者になっても改善されない誤りは何か，といった情報を得ることができる。そして，それらの情報は，言語教育の現場に直接役立つものであり，想定する使用者の習熟度を絞った辞書を編纂する際にも有益な情報となるであろう。

註

本研究の成果の一部は，日本学術振興会科学研究費補助金（特別研究員奨励費（PD 実験））「パターン認識と自然言語処理の技術を用いた習熟度判定」（代表: 小林雄一郎）（2012-2014 年度），立命館大学学内公募型研究推進プログラム（若手研究）「データマイニングによる誤り分析と辞書開発に向けた基礎研究」（代表: 小林雄一郎）（2012 年度）によるものである。

参考文献

Abe, M. (2012). Variation across proficiency levels in L2 spoken English. *Proceedings of TaLC 10* (Online).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of 20th International Conference on Very Large Data Bases*, 487-499.

Gillard, P., & Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In Granger, S. (ed.), *Learner English on computer* (pp. 159-171). London: Longman.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *ICAME Journal,* 28, 31-48.

和泉絵美・内元清貴・井佐原均 (2004). 『日本人 1200 人の英語スピーキングコーパス』 アルク.

Nagata, R., & Nakatani, K. (2010). Evaluating performance of grammatical error detection to maximize learning effect. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING),* 894–900.

Pahl, C., & Donnellan, D. (2002). Data mining for the evaluation of web-based teaching and learning environments. *Proceedings of E-Learn 2002,* 747-752.

Rundell, M. (1998). Recent trends in English pedagogical lexicography. *International Journal of Lexicography,* 11(4), 315-342.

投野由紀夫 (2012). 「学習者コーパス情報を辞典に活かす」 英語コーパス学会東支部課題別シンポジウム「コーパス分析と辞書」 2012 年 3 月 18 日，成城大学.

Wang, F. H., & Shao, H. M. (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining. *Expert Systems with Applications,* 27(3), 365-377.