The NICT JLE Corpus における発達指標の研究
――コレスポンデンス分析によるタグ頻度解析―― *

小林　雄一郎　（法政大学大学院）
kobayashi0721@gmail.com

1.　はじめに

学習者コーパスとは、外国人学習者によって産出された言語データを機械可読な形式で集積したものである [1]。近年、様々な学習者コーパスが編纂され、習熟度レベル別の学習者による発表語彙を多角的に分析することが可能になっている。そのような状況で、どのような語彙・文法・形態の使用（または、誤用）パターンが習熟段階を判定する「発達指標」となり得るのかを特定する組織的研究は、世界的にもまだ少ない。

本研究は、多変量解析を用いて、学習者コーパスにおける発達指標に光を当てるものである。具体的には、CLAWS (C7 tagset) で情報付与したデータにコレスポンデンス分析を実行し、初級者と上級者を分ける品詞・文法項目の使用頻度を明らかにする。

2.　背景

2. 1.　最近の学習者コーパス

初期の学習者コーパス研究の多くは、異なる母語を持つ同レベルの学習者データを比較し、母語転移などの平行方向の分析を目的とするものであった。それに対して、The NICT JLE Corpus [2] は、学習者レベルごとにサブ・コーパス分けされており、発達指標などの垂直方向の分析を可能にする唯一の大規模データである。因みに、中高生の作文データを集めた JEFLL Corpus [3] は、学年でサブ・コーパス分けされているものの、厳密な学習者レベルで分けられている訳ではない。

2. 2.　多変量解析を用いた発達指標研究

この分野における主な先行研究は 2 つある。先ず、Tono [4] は、JEFLL Corpus における品詞タグ連鎖にコレスポンデンス分析を実行し、学習者言語が名詞句中心から動詞句中心へと推移し、その後に前置詞句が発達していく過程を明らかにした。また、Abe [5] は、The NICT JLE Corpus における名詞関連と動詞関連のエラーにコレスポンデンス分析を実行し、初級者には動詞関連のエラーが顕著であるのに対して、上級者には名詞関連のエラーが顕著であることを示した。

これらの研究を参照する限り、学習者の習熟度は、名詞および動詞の産出と密接に関係しているように思われる。

2. 3.　第 1 言語における品詞習得の研究

第 1 言語における品詞習得研究としては、Gentner [6] が挙げられる。これは、英語を含む複数の言語において、幼児が動詞よりも名詞を先に産出し、知覚することを明らかにした。それ以降、第 1 言語における名詞と動詞の習得順序に関して多くの研究が行なわれてきたが、第 2 言語における同様の研究は殆ど見つけられない。

3.　研究目的

本稿の目的は、以下の 3 つの問いに答えることにある。(1) 学習者レベルによって、品詞や文法項目の頻度に差があるのではないか。(2) 品詞タグの頻度を多角的に分析することで、学習者レベルの識別ができるのではないか。(3) 第 2 言語に

---

* Investigating Developmental Criteria in the NICT JLE Corpus through Correspondence Analysis of Word-Class Distribution, Yuichiro KOBAYASHI (Hosei University Graduate School)

おける品詞習得の順序は、第 1 言語の習得順序と相関があるのではないか。

4. データ

4. 1. コーパス

実験に用いるデータは、The NICT JLE Corpus である。このコーパスは、日本人英語学習者約 1,200 人の話し言葉データ（インタビュー形式）を集め、SST レベルと呼ばれる 9 つの習熟度レベルを付与したものである（総語数は約 200 万語）。今回は、学習者発話（約 130 万語）のみを対象とし、総語数の少ないレベル 1 とレベル 2 を 1 つにまとめた（表 1）。

	学習者数	総語数	学習者発話の語数
レベル 1&2	38	39,832	17,008
レベル 3	222	297,764	175,262
レベル 4	482	778,573	510,807
レベル 5	236	437,648	306,264
レベル 6	130	257,621	183,584
レベル 7	58	120,994	87,269
レベル 8	25	56,853	42,888
レベル 9	10	22,385	16,318
合計	1,201	2,011,670	1,339,400

表 1　各レベルのデータ数と総語数

4. 2. 品詞・文法項目タグ付与

本研究で用いる品詞・文法項目タグは、CLAWS (C7 tagset) [7] である。これは、確率論に基づくタガーで、135 種類のタグを持っている。その精度は、一般に 95～97%と言われているが、学習者データの場合はそれほど高くない。そのため、今回用いるデータは、一部手作業で修正してあるものの、いまだにタグの付け間違いが存在する。この点に関しては、今後の課題とする。

5. 方法論

実験に用いる方法論は、多変量解析の 1 つであるコレスポンデンス分析である。この手法は、大量のデータを分類・整理・縮約し、個体間の関係、変数間の関係、個体と変数の関係を多次元空間上に視覚化するものである。また、近年、コーパス言語学の分野でもテクスト類型の手法として頻繁に用いられている。

コレスポンデンス分析に用いる変数は、CLAWS のタグ全てである（コーパス全体の生起頻度が 0 のものは除く）。この解析法は、Tabata [8] が述べているように、(1) コーパス全体を射程に収めることができる、(2) 個体間の文体的な差異が強調され、内容や主題の影響を受けにくい、(3) 任意の変数を選択する研究者の恣意性を排除できる、といった利点を持っている。

6. 結果と考察

コレスポンデンス分析を実行し、最も寄与率の高い第 1 次元 (81.29%) と第 2 次元 (11.59%) の得点を 2 次元散布図に投影したものが、以下の図 1 および図 2 である。

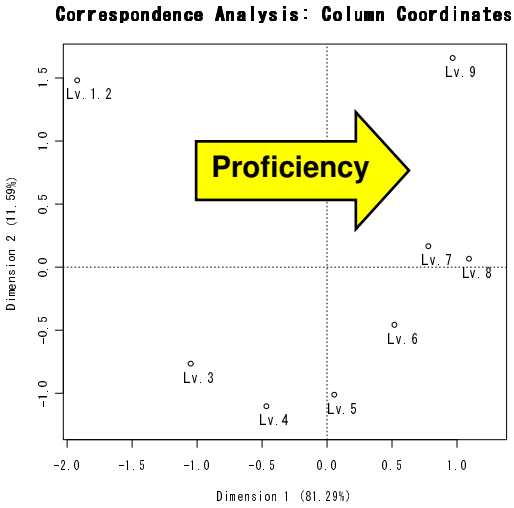

図 1　個体間の相互関係

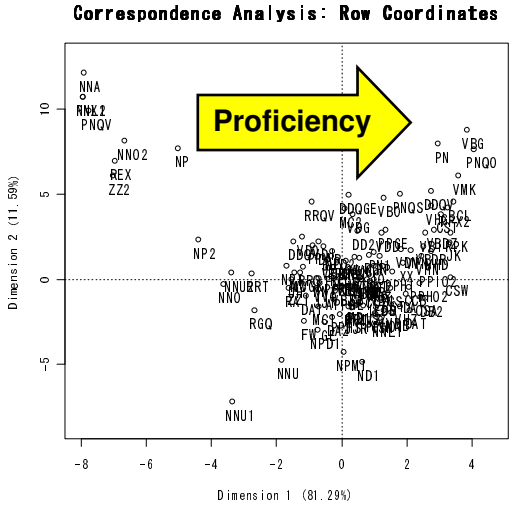

図 2　変数間の相互関係

図中で近接する項目は類似した性質があることを示し、図中の項目間を隔てる距離が大きいほど異質性が高いことを示す。また、2 つの図を照合することにより、個体（SST レベル）と変数（タグ）の関係を読み取ることができる。なお、統計処理には、オープンソースの統計処理言語（環境）である R を用いた。

6. 1.　個体間の相互関係

最も寄与率の高い（各個体間の関係性を最もよく説明する）2 つの次元に、学習者レベルが反映されている。この結果は、品詞・文法項目タグの頻度が学習者レベルを判定する発達指標となり得ることを示している。

6. 2.　変数間の相互関係

図 2 は、学習者がよく使う品詞・文法項目（CLAWS のタグ）の特徴が、SST レベルが上がっていくにつれて変化し、それが徐々に目標言語の母語話者の用法に近づいていくことを示している。この図は、非常に情報量が多いため、様々な分析が可能である。その中で、今回は、図中の左側に名詞関連のタグが分布し、右側に動詞関連のタグが分布している点に注目する。以下の図 3 は、図 2 から名詞関連のタグと動詞関連のタグだけを抜き出して表示したものである。

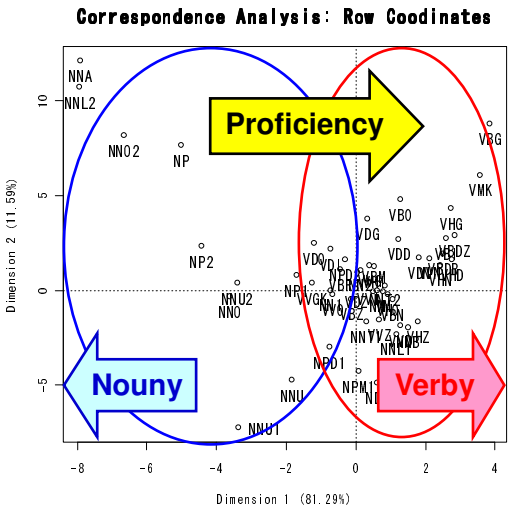

図 3　「名詞中心」から「動詞中心」への移行

図 3 に明らかなように、学習者言語は、SST レベルが上がるにつれて、名詞中心の発話から動詞中心の発話へと移行していく。この結果は、第 2 言語習得においても、第 1 言語の場合と同様に、品詞の習得順序が存在することを意味するのであろうか。何をもって「習得した」と見るかは言語学的に難しい問題ではあるが、少なくとも「初級の学習者は名詞中心の発話をし、上級の学習者は動詞中心の発話をする」と述べることは可能である。

6. 3.　質的分析

言うまでもなく、コーパスに基づく量的分析は、人間による質的分析と相補的な関係になければならない。以下の引用 (1) はレベル 1 の学習者による名詞中心の発話の例である。なお、<A>はインタビュアーの発話を表し、<B>は学習者の発話を表している。

(1) <A>You come by train or bus?</A>
<B>Er. **Train**.</B>
<A>OK. Train. OK, in XXX04, do you live with your family?</A>
<B>Yes.</B>

<A>OK. Please tell me about your family.</A>
<B>Er **wife** and **children** er **child** ka</B>
（学習者発話における名詞の強調は引用者）

初級の学習者は、インタビュアーの質問に対して、基本的に単語（名詞）のみで答える傾向がある。これは、質問の意図は何となく分かっているが、自分の言いたいことを文章にすることができない段階である。これに対して、引用 (2) は、レベル 9 の学習者による動詞中心の発話の例である。

(2) <A>Eh well eh. You know, this is a very safe neighborhood as you know. And er I could come help you tomorrow night.</A>
<B>But tomorrow night **is** a long time. Eh em. **Is** eh I cannot **wait** until tomorrow night em in times of emergency. Em. There might **be** this **is** Japan, but, er in the United States, you might **be** **sued** by me, if something **happens** to me or, can you **reduce** to **rent**?</B>
（学習者発話における動詞の強調は引用者）

上級の学習者の発話には、(話し言葉ということもあって) 一部文法的な誤りはあるものの、動詞を使って長い文章を作ろうという意図が見える。

7. おわりに

本稿では、以下の 3 点が明らかにされた。(1) 学習者レベルによって、品詞や文法項目の頻度には明確な差がある。(2) 品詞タグの頻度は、学習者レベルを識別する発達指標となり得る。(3) 初級の学習者は「名詞中心の発話」をし、上級の学習者は「動詞中心の発話」をする。今後の課題としては、タグが生起するコンテクストの精査、書き言葉での再検証、タグ付け精度の向上、発達指標の絞込みなどが考えられる。

参考文献

[1] Leech, G. (1998) “Preface.” In Granger, S. (ed.) *Learner English on Computer*. London: Longman (pp. xiv-xx).

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

[3] Tono, Y. (2002) *The Role of Learner Corpora in SLA Research and Foreign Language Teaching: The Multiple Comparison.* Unpublished Ph.D Dissertation. Lancaster: Lancaster University.

[4] Tono, Y. (1999) “A Corpus-Based Analysis of Interlanguage Development: Analyzing Part-of-Speech Tag Sequences of EFL Learner Corpora.” In Lewandowska-Tomaszczyk, B. and P. J. Melia. *PALC '99: Practical Applications in Language Corpora.* Frankfurt: Peter Lang (pp. 323-340).

[5] Abe, M. (2004) “A Corpus-Based Analysis of Interlanguage: Errors and English Proficiency Level of Japanese Learners of English.” *Handbook of an International Symposium on Learner Corpora in Asia.* Tokyo: Showa Woman's University (pp. 28-32).

[6] Gentner, D. (1982) “Why Nouns are Learned Before Verbs: Linguistic Relativity versus Natural Partitioning.” In S. A. Kuczaj (ed.) *Language, Thought, and Culture.* Hillsdale: L. Erlbaum Associates (pp. 301-334).

[7] Garside, R., G. Leech and A. McEnery (eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman.

[8] Tabata, T. (2002) “Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution.” In Saito, T., *et al.* (eds.) *English Corpus Linguistics in Japan.* Amsterdam: Rodopi (pp. 165-182).