

「日本人英語学習者『動的』コーパス」の構築

阪上 辰也 (名古屋大学), 森田 光宏 (山形大学), 村尾 玲美 (早稲田大学), 松野 和子 (三重大学)

Abstract

「動的」コーパス＝書き言葉の産出過程

近年、学習者コーパスの量的・質的な分析から、特に、学習者が産出した定型表現（ひとまとまりとして慣習的によく使用されている表現）の量や種類に関する特徴が明らかにされてきている。(Granger, 1998; 杉浦, 2000)。

しかしながら、一連の研究では「頻度のみ」を基準として定型表現を抽出している。本研究の出発点は、頻度のみから定型表現を認定してよいのかという疑問にある。

本研究は、頻度という認定基準に、「**産出速度**」という基準を新たに加え、定型表現の抽出と検討を試みる。この研究目的を達成するため、新たな研究方法として「**動的コーパス**」の構築を行い、産出過程の量的・質的分析を行った。動的コーパスには、既存のコーパスでは記録されていない、英作文の産出過程が記録されている。

Backgrounds & Objectives

① 頻度ベースのコーパス分析でよいのか？

【従来】

頻度のみを基準として定型表現を抽出する
(De Cock et al., 1998; 杉浦, 2000)

【実際】

コーパス中の頻度に基づいて定型表現と認定されたものの中には、ひとつのまとまりとして記憶されていないものがある
(Schmitt et al., 2004)

頻度だけに頼らず、まとまりとして産出されるかどうかを「速度」から観察・検証したらどうか

② 処理単位はどうなっているのか？

Reading	句
Listening	句
Speaking	単語以上
Writing	ほぼ未調査

どの単位で産出されるか？（語？句？節？文？）

Methods & Results

【作文過程リアルタイム記録システム】

- Webベース
 - JavaScript（キーのログ・時間取得）
 - Perl（データ管理）
- ログ：1) 問題番号, 2) 時間, 3) 入力文字列
- Firefox, Safari で動作

得られたデータから

「動的コーパス」として構築

【データ提供者】

日本人英語学習者（大学生）：26名

【収集手順】

- ① 提示された英文を入力し、使い方を学習
- ② 「公共の場での喫煙」というトピックで作文
- ③ アンケート調査（システムの使用感など）

【分析例】

- ① 文字を1つタイプするのにかかる平均速度の算出（1 stroke するのにかかるミリ秒の総計／総文字数）
- ② 作文データから n -gram 表現の抽出
- ③ [期待速度 (平均速度×文字数)] と [観測速度 (実際の n -gram 表現の産出速度)] の比較

【結果】

- “people who”は、期待速度よりも有意に遅く産出。
($t(38) = 2.41, p < .05$)
- 残り6つの表現は、期待速度よりも速く産出される傾向あり（有意差は確認されず）。

- in public ($t(48) = 0.84, p = .41$)
- I think ($t(32) = 0.79, p = .44$)
- smoking in ($t(26) = 1.60, p = .13$)
- many people ($t(24) = 1.11, p = .29$)
- there are ($t(24) = 0.04, p = .97$)
- smoke in ($t(18) = 0.49, p = .63$)

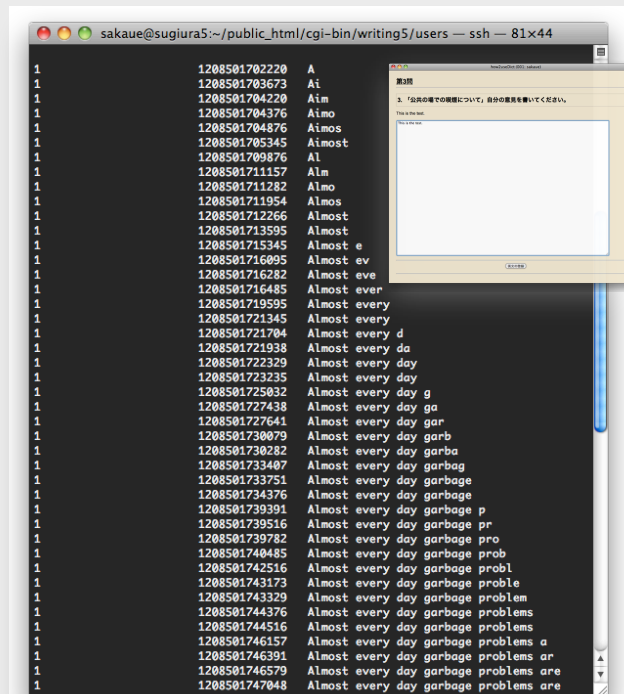


図1 入力画面とログ

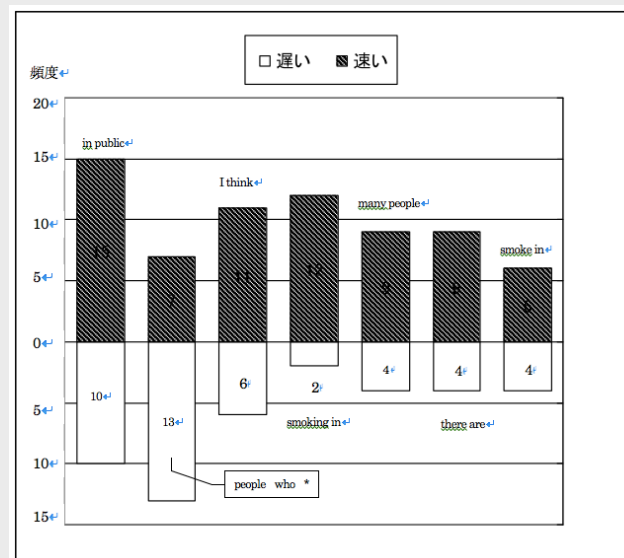


図2 n -gram 表現の期待速度と観測速度の比較

Further Research

- (1) 期待速度より速く(遅く)産出される要因分析
文法・意味・コロケーション etc.
- (2) 3-gram 以上の表現の産出速度計測
 n -gram による定型表現抽出の妥当性の検討
- (3) データ提供者の分類
英語の習熟度、産出過程による分類
- (4) 母語話者との産出単位の比較
- (5) ポーズ位置から見た産出単位の分析
ポーズをどう決めるか
- (6) データ処理の効率化（分析ツール開発）
現在、時間の計算・表現の抽出は手作業
NLP 専門の方の協力を！
- (7) JavaScript による時間記録の信頼性の検証

Conclusions

- (1) 動的コーパスの構築により、書き言葉の産出過程を詳細を捉えられる。
- (2) 量的な分析のみに頼らず、認知的側面を考慮した定型表現の抽出が可能になる。

References

- De Cock, S., S. Granger, G. Leech, and T. McEnery. (1998) “An automated approach to phrasicon of EFL learners.” In S. Granger (ed.) *Learner English on Computer*. London: Longman, pp. 67-79.
- Granger, S. (ed) (1998). *Learner English on Computer*. Addison Wesley Longman.
- Just, M. A. & P. A. Carpenter. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review*, 87, 329-354.
- Kadota, S. (1986). The process of speech production: An analysis of pauses. *NCI Ronso*, 381-398. Tokyo: New Current International.
- 門田修平 (2002). 『英語の書きことばと話しことばはいかに関係しているかー第二言語理解の認知メカニズム』くろしお出版.
- Kohno, M. (1993). Perceptual sense unit and echoic memory. *International Journal of Psycholinguistics*, 9(1), 11-31.
- Schmitt, N., Grandage, S. and Adolphs, S. (2004). “Are corpus-derived recurrent clusters psycholinguistically valid?” In N. Schmitt. (ed.), *Formulaic Sequences: Acquisition, processing and use*. Amsterdam: John Benjamins, pp. 127-151.
- 杉浦正利 (2000) “ n -gram を使用した学習者コーパスにおける共起頻度の分析” 朝尾幸次郎(代表) 『第二言語習得研究のための英語学習者コーパスの構築とその利用』平成9年度-平成11年度科学研究費補助金(基盤研究 (B) (1)) 研究成果報告書, 研究課題番号 09558018, pp. 37-43.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84-119.
- Wray, A. and M. R. Perkins. (2000) “The functions of formulaic language: an integrated model.” *Language & Communication* 20: 1-28.