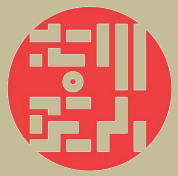
2010年11月13日
日本英語学会第28回大会
(ELSJ28)@日大文理

ワークショップ第7室
「コーパスを用いた構文研究の新展開」
(スチューデント・ワークショップ)

構文研究における コーパス利用の理論

方法論を超えて

吉川 正人 machayoshikawa@dream.com
慶應義塾大学大学院/日本学術振興会特別研究員



I.　はじめに

コーパスと構文研究

- 意外とコーパスベースの構文研究は少ない?
  - Construction Grammar Bibliography より
    - http://www.constructiongrammar.org/ (2010年9月27日更新分)
    - “corpus” とタイトルに入っている文献= 19/347 ≈ 5.5 %
      - うち 3分の2 近く (12) がStefan Th. Gries (ら) によるもの
    - Boas (2003) などタイトルに現れていないコーパス研究もある
    - しかし: 多くが単に「コーパスを使った」だけの研究では!?
      - 少なくともHundt (2001), Boas (2003)はそう
- 実情
  - コーパスを用いた構文研究の方法論が未成熟

➔ 方法論を下支えする分析の「理論」の不在が原因?

本WS及び本発表の目的

- コーパスを用いた構文研究を新たなステージへ
  - コーパスを用いる意義を再検討
    - コーパス = 使用実態のサンプル
    - ~~集めたデータを分析する~~
      - ➔ 手続き *P* に従いデータ *D* を解析し結果 *P*(*D*) を評価
  - ただし: あくまで考えて行くべき「道筋」の提案
- 概要
  - 理論的前提と用いる方法との対応を熟慮する
  - 再現率と適合率を見積もる
  - 「検索と分析」の二重性を念頭に置く

発表の構成

- 2節: コーパス調査一般に関わる問題
  - 手続きの統制と結果の評価
  - 再現率と適合率
  - 入力としてのコーパス・出力としてのコーパス
- 3節: コーパスを用いた構文研究に関わる問題
  - 表層にない情報をどう検索するか
  - 検索結果を評価する手順
- 4節: 結語

2.　コーパス調査の方法

「方法論」を超える

- コーパスを「使う」だけでは限界がある
  - コーパスに「ない」ものに関しては何も言えない
  - 理論的前提 & それに裏打ちされた手続きが必要
- コーパス言語学 = 方法論? (Leech 1992: 105-107)
  - それでいいのか?
  - 「コーパスを利用する」
    ➔「コーパスをある手続きP で解析する」
    - コーパス言語学 = この手続きの正当性を保証する理論体系
  - 理論的な下支えがあって初めて結果の評価が可能

コーパスを用いる意義

- 「言える/言えない」以上のことが言える
  - 許容範囲 = 境界条件の特定なら作例の方が有益
  - “all or nothing” より “more or less” の議論が得意
    - 前提:コーパス上の分布特性がヒトの容認性判断と相関を持つ
      - 主として「入力」としてのコーパスの議論に重要 (後述)
  - 表現それ自体もそうだがその生起文脈の特定が容易
    - 作例では「文脈」や「状況」の特定は困難
      - 主として「出力」としてのコーパスの議論に重要 (後述)
- 従って
  - 「どれくらい」「どんな時に」言えるかを特定できる

再現率と適合率 [1]

- コーパス調査の基本手法 = 検索
  - 最も良い検索
    = 1) 探したいもの全てにヒットし
      2) 探したいもの以外に全くヒットしない
  - 1 の達成度 = 再現率 (recall)
    - 再現率 = 見つかった検索対象の総数 *R* / 検索対象の総数 *C*
  - 2 の達成度 = 適合率 (precision)
    - 適合率 = 見つかった検索対象の総数 *R* / 検索結果の総数 *N*

再現率と適合率 [2]

- だが: 両者は一般に反比例する
  - 目的にあった検索を行う必要アリ
  - 重要な点
    - C を正確に知ることはほぼ困難 ➔ 近似値を推定する
    - この推定には理論が不可欠 (Cf. 吉川 2010b)
      - E.g., 選好項構造 (Du Bois et al. 2003) の活用・典型パターンの特定
- 検索結果の(人手)クリーニングの効果
  - クリーニングによって適合率は上昇する

    ➔ が: 再現率は上がりようがない
  - ある程度の再現率を保証する必要がある

コーパスの位置づけ

- コーパス = 言語の使用実態のサンプル
  - これには実は二面性がある
    - 1. 誰かが「使った」データの集まり = コーパス as 出力
      2. 誰かが「見聞きした」データの集まり = コーパス as 入力
  - 「認知」や「学習」と結び付けるなら熟慮すべき区別
    - Cf. 用法基盤モデル (Usage-based Model)
  - 「入力」と考える → 頻度分布の有効利用
    - Cf. 分布バイアス仮説 (Shirai & Anderson 1995)
  - 「出力」と考える →「実態調査」的な手法
    - 伊澤発表も参照

3.　コーパスで構文を切る

二つのアプローチ

- コーパス基盤 (corpus-based)
  - 方法
    - ある構文 C (e.g., 二重目的語構文)をコーパスから(自動)収集
    - 得られたデータの性質を分析
  - 従来の多くの研究がこれ
- コーパス駆動 (corpus-driven)
  - 方法
    - なんらかのアルゴリズムに従いコーパスデータを解析
    - 得られたデータから何らかの構文(群) C を特定 or 発見
  - これはほとんどない
    - Cf. Hunston & Francis 2000; Sinclair & Mauranen 2006

仮説生成と仮説検証のサイクル

- 「純粋」なコーパス駆動研究の難しさ
  - そもそもコーパスを見てみないことには何も言えない
    - ▶ 前提となる理論的研究も「データを見て」いる
  - 頑健な分析にはコーパス基盤の「仮説生成」が不可欠
  - 「仮説生成」と「仮説検証」のサイクルを認識すべき
- 構文研究には特に重要
  - 検索方法 Q で構文 C の収集可能かどうか

    ➔ 試行錯誤によってのみ判明

コーパスと構文の相性

☼ 構文情報はコーパス上にはない

- 統語解析済みコーパスでさえ情報は限定的
  - ▶ 二重目的語構文はいいが結果構文・使役移動構文等はダメ
- 多くのコーパスは品詞タグ & レンマ情報のみ

- 「見えない」情報をうまくすくい取る必要がある

➔ 情報を「見える化」する仮説の生成が不可欠

- ▶ 中村発表も参照

手続きの二重性

☼ コーパス基盤構文研究における手続きの二重性

- 1. 対象となる構文 C の(自動)収集 = 検索
  2. 得られた構文 C の事例群の分析 = 分析
  - ▶ コーパス駆動なら基本的には1のみ (+ 結果の評価)
- 重要: 両者は全く別のプロセス
  - ▶ 「見えない情報をどう見つけるか」= あくまで検索の問題

- 区別して考え独立に評価する必要がある

Collostructional Analysis

- 体系化された構文分析手法の代表例
  - 主にAnatol StefanowitschとStefan Th. Griesによる
    - Stefanowitsch & Gries (2003)など文献は多数
  - 概要 (collexeme analysis):
    - 構文 c とその変項を実現する語との共起頻度を分割表で表現
    - 分割表に基づきフィッシャーの正確検定を実施
    - 得られた p 値 (の負の対数)を連想強度 としてランク付け
      - 期待値と実測値の差の正負も活用
  - 前提
    - 計算した統計量がヒトの認知処理と対応している
      = ヒトの統計処理能力 & 用法基盤モデルを想定

再び、手続きの二重性

- Collostructional analysis における二つの手続き
  - 手続き I は(主に)統語解析済みコーパスの使用で対応
    - ICE-GBなどがよく使われる
  - 従って: 重要なのは手続き2
- だが: 手続きI はどうでもいいわけではない
  - 再現率と適合率の問題・仮説生成プロセス
  - 手続きI の相対的非重要性の起源
    = 精度がコーパスの構造・検索ツールの性能に依存
    - 自作コーパス・解析ツールの使用で重要度は上昇

プログラミングの勧め

- 与えられたツールの使用には限界がある

- 自前で処理スクリプトを用意すれば …
  - 解析済みのコーパスがローカルにあれば直接処理可能
  - タガー・パーザーが手に入れば解析も自前で実行可能
- Python の勧め
  - フリーで言語処理用のツール群が充実
    - NLTK (Natural Language Toolkit)
      - 10月9日の英語コーパス学会のワークショップでも紹介された
  - 発表者は使用約2年だがそれなりに熟達 = 簡易

4.　結語

まとめ

- 一般論
  - 用いる方法の理論的意義
  - 検索の再現率・適合率
    - 原則再現率を上げるべき
  - 入力/出力どちらとしてコーパスを見るのか
- 構文研究
  - コーパス基盤/駆動・仮説生成/検証の別
    - 構文とコーパスの相性を考えると仮説生成は必須
  - 手続きの二重性 = 検索 + 分析

6.　謝辞と参考文献

謝辞 (五十音順)

☼以下の方々にこの場を借りて謝意を表します

- 黒田 航氏　(京都工芸繊維大学非常勤講師)
- 長谷部 陽一郎氏　(同志社大学)
- 日高 昇平氏　(北陸先端大学)

以下ワークショップ参加メンバー

- 中村 文紀氏　(慶應義塾大学大学院)
- 伊澤 宜仁氏　(慶應義塾大学大学院)
- 野中 大輔氏　(慶應義塾大学大学院)

参考文献

- Du Bois, J., Kumpf, L., & Ashby, W. 2003. *Preferred argument structure: Grammar as architecture for function*. Amsterdam: John Benjamins.
- Hunston, S., & Francis, G. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
- Langacker, R. 1987. *Foundations of cognitive grammar Vol. 1: Theoretical prerequisites*. Stanford: Stanford University Press.
- Leech, G. 1992. Corpora and theories of linguistic performance. In Startvik, J. (ed.) *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 105-122). Berlin: Mouton de Gruyter.
- Shirai, Y., & Anderson, R. 1995. The acquisition of tense-aspect morphology: A prototype account. Language, 71, 743-762.
- Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J., & Mauranen, A. 2006. *Linear unit grammar: Integrating speech and writing*. Amsterdam: John Benjamins.
- Stefanowitsch, A., & Gries, S. Th. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*, 209-243.
- 吉川正人. 2010a. 「語」を越えた単位に基づくコーパス分析に向けて: パターンラティスモデル(PLM) とその有用性. 『藝文研究』 98, 221–207.
- 吉川 正人. 2010b. 「構文の多義」再考: 事例基盤構文理論に向けて. 『日本認知言語学会論文集』, 10, 449-459.

ご清聴有難うございました。