日本語コーパスを用いた複文研究の可能性

丸山岳彦（国立国語研究所）

国立国語研究所共同研究プロジェクト
「複文構文の意味の研究」

2011/2/19 神戸「ユニティ」

1

発表の流れ

2

- コーパス日本語学の射程
  - コーパスを使うことの意義
- 日本語コーパスの現状
  - 世界のコーパス
  - 日本のコーパス
- コーパスを用いた複文研究の可能性
  - コーパスの検索と用例の抽出、集計
  - 検索・分析例：「並列節」と「条件節」

コーパス日本語学の射程

3

- 「コーパス日本語学」：コーパスを用いて日本語を研究する方法論

コーパス	研究領域
現代日本語　書き言葉 近代日本語　話し言葉 近世日本語 中世日本語 古代日本語	音声・音韻 語彙 文法 意味 談話構造 …

コーパスを使うことの意義(1)

4

- 内省できない「行動」の可視化

問題）言いよどむ際に現れるフィラーの形式を、思いつくだけ挙げなさい。

えーと、

問題）自分が最も多く使用するフィラーの形式を考えなさい。また、日本人が最も多く使用するフィラーの形式を考えなさい。

コーパスを使うことの意義(2)

- 定量的な分析に基づく定性的な分析

問題）「ハシゴにのぼる」と「ハシゴをのぼる」の意味の違いを考えなさい。

問題）「登る」「上る」「昇る」「のぼる」は、それぞれ「～を」「～に」にどのような名詞を取るか、考えなさい。

コーパスを使うことの意義(3)

- 大量の用例を容易に取得できること
- 条件に応じて柔軟な検索ができること
- 内省では気づかない言語現象を発見できる（可能性がある）こと
- 追試ができること
- メディア・ジャンル・使用場面の違いを考慮した言語現象の分析ができること

	のぼる		登る		上る		昇る	
～に	N 人	52	山	87	N 人	70	天	18
	N 円	30	木	35	N 円	43	脚立	7
	N %	20	上	13	N %	35	上	6
	話題	20	丘	8	N 件	20	位	4
	数	16	頂上	6	話題	19	N 階	3
～を	階段	44	石段	32	階段	111	階段	29
	坂	11	階段	29	坂	21	スロープ	2
	坂道	11	坂	21	石段	12	梯子	2
	丘	6	斜面	18	坂道	11	煙突	1
	石段	4	坂道	18	山道	6	傾斜	1

『現代日本語書き言葉均衡コーパス』3,500万語に現れた「のぼる」の集計（上位5位）

世界のコーパス

- 計画的なコーパスの構築は1950年代から

1959年 英 The Survey of English Usage （100万語）
1964年 米 Brown Corpus（100万語）
1991年 英 Bank of English (BOE) （～5.5億語～）
1994年 英 British National Corpus (BNC) （1億語）
2006年 台湾 Sinica Corpus（～10億語）
2007年 韓国 Sejong Corpus（1億語）

- 語彙表の作成、記述文法書の編纂
- コロケーション、コリゲーション

9

日本のコーパス(1)

- 国立国語研究所による語彙調査
  - 1953年 『婦人雑誌の用語』
  - 1957-1958年『総合雑誌の用語』
  - 1962-1964年『現代雑誌九十種の用語・用字』
  - 1970-1973年『電子計算機による新聞の語彙調査』
  - 1983-1984年『高等学校教科書の語彙調査』
  - 1986-1987年『中学校教科書の語彙調査』
  - 1995-1999年『テレビ放送の語彙調査』
  - 1994-2005年『現代雑誌200万字言語調査』
- 語彙表の作成、教育基本語彙の選定

→ 調査資料（＝コーパス）は非公開

10

日本のコーパス(2)

- 『日本語話し言葉コーパス（CSJ）』（2004年）
  約661時間、約752万語の自発音声コーパス。
- 『太陽コーパス』（2005年）
  総合雑誌『太陽』（1895年～1925年発行）のテキスト。1,450万字。
- 『現代日本語書き言葉均衡コーパス（BCCWJ）』
  さまざまなメディアの書き言葉を収録した1億語のバランストコーパス。2011年度に一般公開予定。

11

コーパスの検索

- コーパスの検索と用例の抽出
- デモ(1) BCCWJの本文テキストの検索
  - 全文検索ソフト「ひまわり」
- デモ(2) CSJの形態論情報の検索
  - 転記テキスト、形態論情報

12

コーパスを用いた複文研究(1)

- 話し言葉に現れる並列節・条件節の分析
- 対象データ：『日本語話し言葉コーパス』
  - 学会講演（改まったスタイルの発話）
    987講演、274.4時間、3,344,616語
  - 模擬講演（くだけたスタイルの発話）
    1,715講演、329.9時間、3,657,277語

13

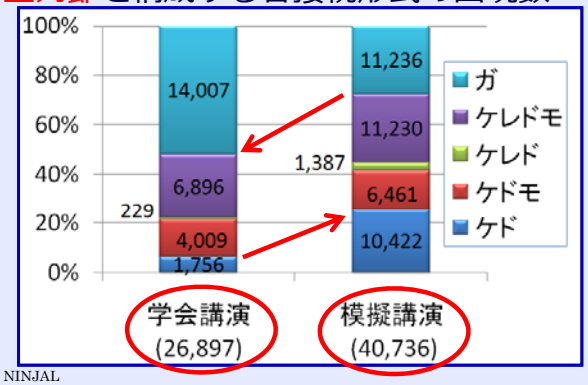
コーパスを用いた複文研究(2)

- 調査対象：
  - 並列節（ガ、ケレドモ、ケレド、ケドモ、ケド）
  - 条件節（レバ、タラ、ナラ、ト）
- 調査項目
  - 出現数の分布は学会講演と模擬講演で異なるか？
  - 前接する述語句の丁寧度に違いはあるか？

14

コーパスを用いた複文研究(3)

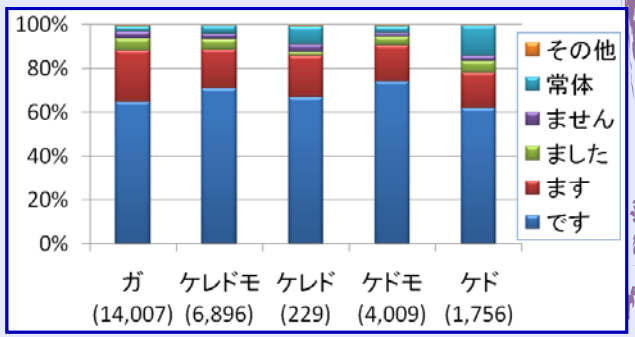
- 並列節を構成する各接続形式の出現数



15

コーパスを用いた複文研究(4)

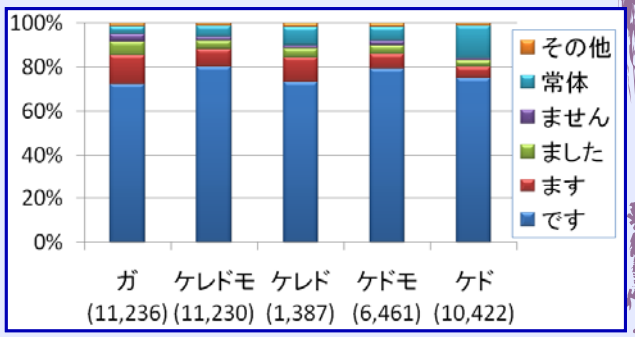
- 各並列節に前接する形式 （学会講演）



16

コーパスを用いた複文研究(5)

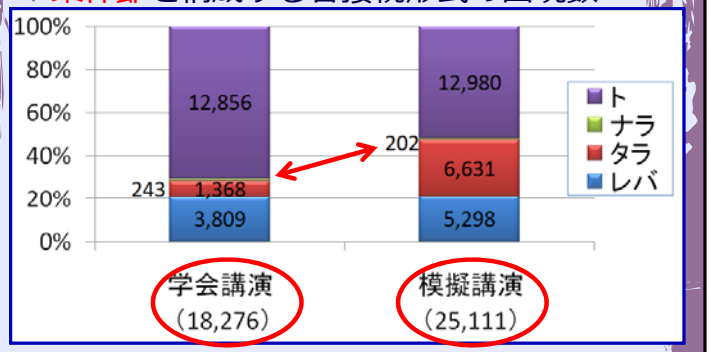
- 各並列節に前接する形式 （模擬講演）



17

コーパスを用いた複文研究(6)

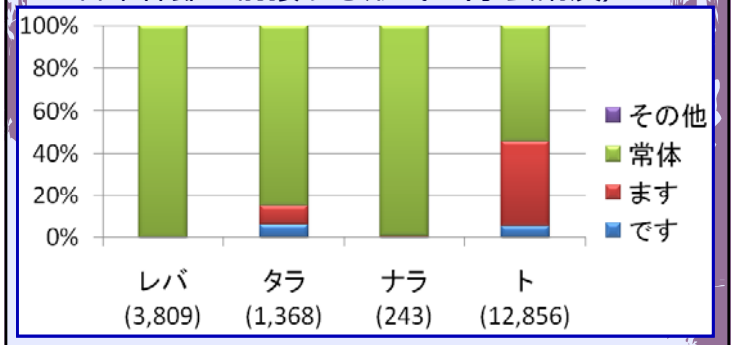
- 条件節を構成する各接続形式の出現数



18

コーパスを用いた複文研究(7)

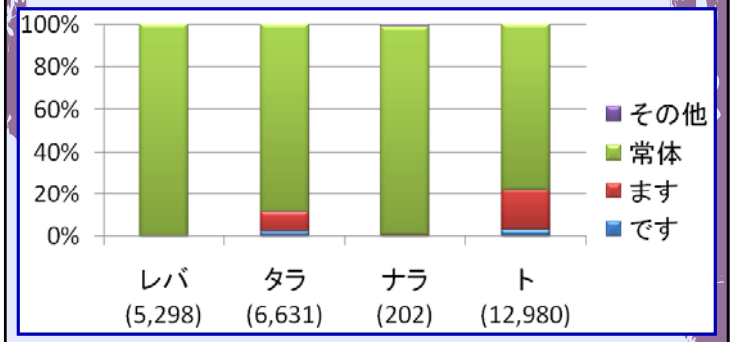
- 各条件節に前接する形式 （学会講演）



19

コーパスを用いた複文研究(8)

- 各条件節に前接する形式 （模擬講演）



20

コーパスを用いた複文研究(9)

- 三尾砂(1942)の「丁寧化百分率」
  - 文末がデスマス体の場合に、従属節が丁寧化する比率を調査（戯曲 ≒ 話し言葉）

が	94.5%
けれど	86%
と	7.3%
ら	6%

今回の調査結果とほぼ同一

定量的な観点を備えた、実証的な文法記述

21

コーパスを用いた複文研究(10)

- コーパスを使う上で留意すべきこと
  - 意味の違いを考慮した検索ができない
    - カラ節：「事態の原因・理由」と「判断の根拠」
    - テ節：「付帯状況」「継起」「因果関係」「並列」
    - 連体節：「内の関係」と「外の関係」
  - → 文法的なアノテーション（情報付与）が必要
  - 集計処理にある程度の技術が必要
  - そのコーパスを理解することが必要
  - → 集計結果が何を意味するのか？

22

参考URL

- 『日本語話し言葉コーパス（CSJ）』
  http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/
  → 音声サンプル、報告書の全文などが閲覧可。
- 『現代日本語書き言葉均衡コーパス（BCCWJ）』
  http://www.ninjal.ac.jp/kotonoha/
  → BCCWJの設計などを簡単に紹介。
- BCCWJ検索デモンストレーション
  http://www.kotonoha.gr.jp/demo/
  → 構築中のBCCWJのうち、作業が終了したデータを公開。簡単な検索が可能。

23

参考文献

- 国立国語研究所 (2006). 『国立国語研究所報告書 124 日本語話し言葉コーパスの構築法』. 国立国語研究所.
- 丸山岳彦 (2009). 「日本語コーパスの現状」. 『国文学 解釈と鑑賞』平成21年1月号（特集 日本語研究とコーパス）, 122-130. 至文堂.
- 丸山岳彦・田野村忠温 (2007). 「コーパス日本語学の射程」. 『日本語科学』. 22, 5-12. 国書刊行会.
- 三尾砂(1942) 『話言葉の文法（言葉遣篇）』帝国教育会出版部