様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月　５日現在

機関番号：１３９０１
研究種目：基盤研究(C)
研究期間：2010　～　2012
課題番号：22520494
研究課題名（和文）　言語研究資料としてのコーパスデータの客観性と信頼性に関する考察
研究課題名（英文）　On the Objectivity and Reliability of Corpus Data for Linguistic Research
研究代表者
大名　力（ Tsutomu OHNA ）
名古屋大学・国際開発研究科・教授
研究者番号：00233205

研究成果の概要（和文）：“ユーザーフレンドリー”な環境の普及によりコーパス，手法等のブラックボックス化が進む現状を踏まえ，コーパス研究における基礎データの信頼性等に関して検討を行った。実際の研究を基に基礎データとして示される頻度数に問題がないか検証し，問題があるものについては誤りが生じた原因について考察し，分類を行った。また，心的実在物として文法を捉える立場から，言語研究資料としてのコーパスデータの性質について検討を行った。さらに，コーパス（データ）の代表性，対象の選択・提示方法，統計値の解釈と検定，“コロケーション”という用語の多義性，語の共起関係に関する計量的指標（主として t-score と MI-score）の信頼性と妥当性等について考察を行い，それらの成果を論文，書籍等で公開した。

研究成果の概要（英文）：With the sizes of corpora increasingly large and with the development and spread of “user-friendly” environments, corpora are now widely used in linguistic research, but these “user-friendly” environments, in spite of their usefulness, also make corpora and corpus studies a kind of “black box”; as a result, users often pay attention only to the output of software while disregarding the input and the process and not examining whether the output can be interpreted as appropriate data for their studies. For the improvement of corpus-based research, I conducted the following surveys: i) I first examined basic data shown in several articles—when errors were detected, I inferred the processes and the causes of the errors, then classified them; ii) from the standpoint of viewing a grammar as a part of the internal state of a speaker, I examined what kinds of information about what aspects of the grammar can be obtained from corpora publicly available now; and iii) I also examined, among other issues, problems concerning representativeness of corpora and corpus data, the reliability and validity of statistical scores such as t-score and MI, and the ambiguity of the term “collocation.”

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	500,000	150,000	650,000
2011 年度	500,000	150,000	650,000
2012 年度	400,000	120,000	520,000
総　計	1,400,000	420,000	1,820,000

研究分野：英語学，言語学，統語論
科研費の分科・細目：言語学・英語学
キーワード：言語学，コーパス，文法，語法

１．研究開始当初の背景

一般に利用可能なコーパスの増加，コーパスの大規模化，所謂“ユーザーフレンドリー”な環境の整備により，特別な訓練を経ずとも大規模コーパスが利用できるようになり，言語研究におけるコーパスの利用は拡大してはいるが，研究の基礎となるコーパスから得られたデータの客観性と信頼性に関しては暗黙の前提となっていることが多く，体系的な調査，考察はもちろんのこと，個々の研究のレベルにおいても，問題の検討がなされることはあまりない。また，“ユーザーフレンドリー”な環境の普及により，コーパス自体がブラックボックス化するのみでなく，手法や方法論もブラックボックス化していく傾向があり，研究者自身もプログラムを利用して行っている処理の内容について意識しにくく，検証の必要性自体が認識されにくい状況が生じている。

２．研究の目的

上記の状況を踏まえ，コーパスからのデータ抽出方法の適切性，統計処理，特に“コロケーション”研究で用いられる共起性の指標に焦点を当て，その内容と利用法・解釈の妥当性を検証し，コーパスデータの客観性・信頼性の確保のために必要な条件を検討する。また，概念の明示化，モデル化，体系的な知識・技術の共有があまり進んでいないため，“共通言語”に乏しく，現実に様々なレベルで問題が生じていても研究者間で明示的かつ体系的に問題を検討することが難しい状況を踏まえ，言語研究資料としてのコーパスデータの性質やその利用法等を批判的に検討し，問題点やその原因を明らかにすることで，今後の議論のための叩き台を提供することも目的とする。

３．研究の方法

論文や口頭による研究発表では，紙幅や時間の都合等で，処理内容が明示的に示されていなかったり，また，データも一部，しかも数値データしか示されておらず，具体的な処理方法がはっきりしないことが少なくない。したがって，論文における記述によって直接内容を検討することが困難なため，論文等で使用されたものと同じコーパスを用いて，基礎となる頻度データが正しいかどうか確認する。問題がある場合は，試行を繰り返すことで同様の値が得られる条件を探り，処理方法を推定し，処理全体のプロセスを踏まえ，誤りが生じた原因を同定する。

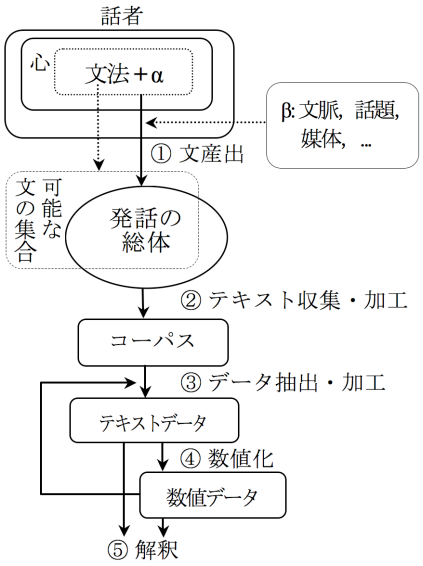

図１ データの抽出と解釈の過程

統計値のうち，共起性の指標については，最初に言語研究への利用を提案した論文まで遡り，計算式，利用方法等を確認する。結果のみを示し計算式を示していない論文が多く，また，それらの指標の計算を自動で行ってくれるプログラムでも利用者にわかる形で計算式を示していないものも多いため，論文，プログラムで使用されている式の内容を，示された／得られた数値から逆算し推定するとともに，利用の仕方に問題がないかを検討する。

また，コーパスからデータを抽出し情報を読み取る過程，それに関わる事項，また，その過程において暗黙のうちに仮定されている前提などを明示化することで，問題がどの段階で生じたものか，問題の原因は何であるかを，特定しやすくする。

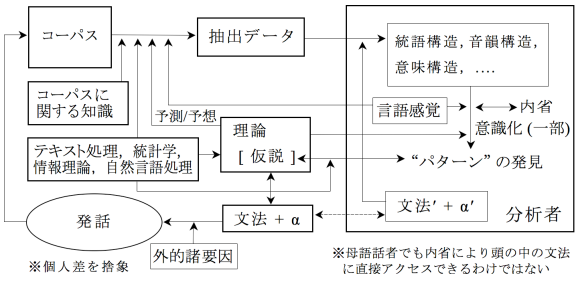

図２

４．研究成果

初年度の２２年度は，対象・処理内容の明示化と基礎データの客観性・信頼性の問題を中心に作業を進めた。定性的分析だけでなく

定量的分析が可能であることをコーパス利用の利点とする研究者は多いが，前述のブラックボックス化によりコーパスデータによる定量的分析には問題も生じやすくなっているため，実際の研究を基に，基礎データとして示される頻度数に問題がないかを検証し，問題の原因および問題を回避するための方策について考察を行った。その成果の一部は，英語コーパス学会第 35 回大会招待講演「コーパス検索で注意すべきこと―基礎データの信頼性向上のために―」，および，日本ドイツ語情報処理学会招待講演「コーパス検索の落とし穴」において発表した。

また，言語研究資料としてのコーパスデータの性質について，心的実在物として文法を捉える立場から，内省データ，実験データなど，他の種類のデータとの関係も考慮に入れながら検討し（図 3 参照），その成果の一部を，日本英語学会第 28 回大会のシンポジウム「文法研究資料としてのコーパスデータの批判的検討」において発表した。

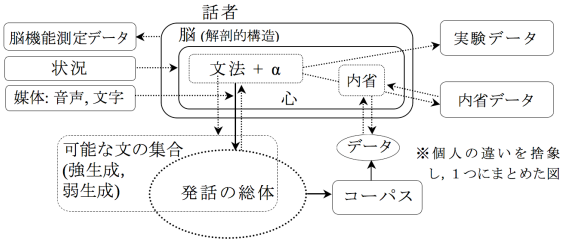

図 3　文法と各種データとの関係

このシンポジウムでは，司会者として現状を整理し，批判的検討を行う必要性を指摘するとともに，講師の１人として，「コーパスから得やすい情報、得にくい情報―統語論、構文研究を中心に」という題目で，文法研究資料としてのコーパスデータの特徴を整理し，統語論，意味論，語用論などの分野に関わる具体的な構文等を示しながら，コーパスから得やすい情報，得にくい情報について検討した。

２３年度は，「統計値の意味とその扱い」・「表現のバリエーションと対象の選択」の問題を中心に検討を行った。具体的には，コーパスとコーパスデータの代表性，コーパス内・サブコーパス間での偏り，対象の選択・提示方法（分類のパラドックス），統計値の解釈と検定，“コロケーション”・“連想関係”等の用語の多義性から生じる問題，連想関係に関する計量的指標（主として t-score と MI-score）の信頼性と妥当性について考察を行い，その成果の一部を「MI-score，t-score と“コロケーション”」という題目で英語コーパス学会第 37 回大会において発表した。この発表では，同じ名称の指標でも研究者やプログラムにより使用している式が異なることを報告し，MI-score, t-score に関して次の 8 点を中心に問題の検討・整理を行った:(a)何に関するものか(連想関係/結び付き/共起(性，関係)/コロケーション(性);(b)(a)の何を計るものか(強度か確信度か);(c)比較の対象は何か(単語の分布のモデルは何か);(d)計算式;(e)期待値の算出方法(期待値はスパンの大きさに応じたものか);(f)“基準値”とその根拠;(g)スコア採択の条件(総語数,各語の頻度,共起頻度,など)とその意味; (h)スコアの利用法とその妥当性。

(d)のスコアの計算式に関しては，1990 年代初頭の Church らの研究に遡り,MI-score, t-score の計算式を確認した上で，他の研究者の論文で示された数値や Bank of English (telnet ベースの旧いタイプ），WordBanks*Online*（ウェブベースの新しいタイプ）などの検索プログラムで示される数値を基に，逆算により使用されている計算式を推定したところ，次の A, Bi, Biia, Biib の４タイプに分類することができることが明らかになった。

A. 連続する x, y のみを対象
B. 指定語数内の x, y を対象
   i. 位置ごとに頻度を集計し MI-score を計算
   ii. 範囲内の位置に現れるものを合計し MI-score を計算
      a. スパン 1 の期待値を使用
      b. スパンに応じた期待値を使用

同じプログラムでも，どのメニューのものかにより，使用されている計算式が異なるものもあった。

計算式が異なれば，同じ数値であっても意味が異なることになる。例えば，Church and Hanks (1990)と Hunston (2002)では，どちらも目安／基準値として MI-score 3 という数値が示されているが，期待値の算出方法に違いがあるため，比較するには換算が必要となる。しかし，上記研究発表に対する参加者の反応からも，こういった情報が研究者の間で共有されていないことは明らかであり，今後，情報の明示化と共有を進めていく必要がある。

どの範囲の表現を対象とするかによって，出現頻度などの基礎的なデータの値も変わってしまうため，定量的分析においては，表

現のバリエーションに配慮することは重要なことであるが，従来の語法文法研究で行われている分析に比べ，コーパスを利用した定量分析では表現のバリエーションへの配慮が不十分と思われるケースが少なくないため，具体例を取り上げ，問題点の整理を行った。また，コーパスデータを処理する際，便宜的に言語学的分類の代わりに表記上の形式的分類を用いることがあるが(例えば，文頭にある副詞を文副詞として扱ったり，直前にコンマが置かれている関係節を非制限用法の関係節として扱ったりするなど)，そのような処理方法に問題はないのか，さらに，コーパスの構造と言語変種の分離（付帯情報・タグ・コーディング，異質な言語変種の混在，メタな言語使用）の問題等についても検討を行った。その結果の一部を整理し，大きく８つに分類し（A. 対象コーパスの選定，B. 検索対象の選定，C. 検索方法，D. 選別方法，E. 分類方法，F. 統計処理，G. 言語学的意義，H. コーパスの規模に対する手法の妥当性)，コーパスを利用する際のチェックポイントとして，堀正広編『これからのコロケーション研究』(出版は 24 年度) の第７章「コーパス利用の落とし穴」でやや詳しく解説した。

２４年度は，前年度に引き続き，“コロケーション”研究でよく用いられる語と語の共起性の指標のうち t-score と MI-score を取り上げ，スコアの信頼性と妥当性についてさらに詳しく検討した。その成果は「“コロケーション”と共起性の指標の信頼性と妥当性について」(*Ex Oriente* vol. 19, pp. 25-52, 大阪大学言語社会学会，招待論文)で公表した。

２４年度は最終年度でもあるため，３年間の研究全体のまとめを行った。特に，心的実在物としての文法（I 言語，I-language）を研究する立場から，コーパスデータの文法研究資料としての性質，利用に際しての注意点について体系的に検討し，それまでの研究成果を有機的に統合することを試みた。また，これまでの研究を踏まえ，コーパスデータの客観性・信頼性を高めるための具体的な方策として，チェックポイントのリスト化・共有，処理内容の明示化，仮説・前提の明示化，複数のコーパスでの検証，スコアの計算式・基準値の根拠等の明示化・検討，研究者間での情報の共有などを進めることの重要性・必要性を具体例を基に検討した。これらの研究の成果については，２５年度以降に公開していく予定である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
①大名力，「“コロケーション”と共起性の指標の信頼性と妥当性について」*Ex Oriente* vol. 19, pp. 25-52. 大阪大学言語社会学会，2012，［招待論文，査読なし］

〔学会発表〕（計４件）
①大名力，「MI-score，t-score と“コロケーション”」，英語コーパス学会第 37 回大会，2011 年 10 月 1 日，京都外国語大学
②大名力，「コーパス検索の落とし穴」，日本ドイツ語情報処理学会 招待講演，2010 年 11 月 28 日，愛知県立大学長久手キャンパス
③大名力，「コーパスから得やすい情報、得にくい情報―統語論、構文研究を中心に」，日本英語学会第 28 回大会 シンポジウム「文法研究資料としてのコーパスデータの批判的検討」，2010 年 11 月 14 日，日本大学文理学部
④大名力，「コーパス検索で注意すべきこと―基礎データの信頼性向上のために―」，英語コーパス学会第 35 回大会招待講演，2010 年 4 月 24 日，兵庫県立大学（神戸学園都市キャンパス）

〔図書〕（計４件）
①堀正広，大名力，他 5 名，『これからのコロケーション研究』，ひつじ書房，2012，担当個所「コーパス利用の落とし穴」pp. 227-264.
②大名力，『言語研究のための正規表現によるコーパス検索』ひつじ書房，2012，216 pp.
③大津由紀雄，大名力，他 17 名，『学習英文法を見直したい』研究社，2012，担当個所「コーパス研究と学習英文法」pp. 256-266.
④藤村逸子，滝沢直宏，大名力，他８名，『言語研究の技法–データの収集と分析』，ひつじ書房，2011，担当個所「言語研究のためのテキスト処理の基礎知識」pp. 259-278，「表計算ソフト、正規表現によるテキスト処理」pp. 279-300.

６．研究組織
(1)研究代表者
大名力（ Tsutomu OHNA ）
名古屋大学・大学院国際開発研究科・教授
研究者番号：00233205

(2)研究分担者なし

(3)連携研究者なし