『日本語話し言葉コーパス』に現れた挿入構造の分析

丸山 岳彦 †‡　　高梨 克也 *　　内元 清貴 *

† ATR 音声言語コミュニケーション研究所　‡ 独立行政法人 国立国語研究所

* 独立行政法人 通信総合研究所

1　導入

話しことばを定量的な観点から観察・記述する研究は，1950 年代より国立国語研究所において活発に行なわれた．特に「話しことばの文型」と題された一連の研究では，実際に録音した話しことばをデータベース化し，文の認定，表現意図，構文，イントネーション，総合的文型などが詳細に記述された [1, 2]．こうした研究は「言語生活」と呼ばれる研究分野の大きな一翼を担ったが，生成文法に代表される理論的研究や日本語教育からの要請を受けた記述的文法研究が盛んになったことにより，1970 年代以降は衰退してしまった．

『日本語話し言葉コーパス (Corpus of Spontaneous Japanese; 以下 CSJ と記す)』は，「学会講演」「模擬講演」を中心とした日本語の自発的な発話 (650 時間，700 万語) が書き起こされ，形態素情報，イントネーションラベル，係り受け構造，談話構造など，さまざまな研究用情報が付与された，大規模な自発音声コーパスである [3]．このような話しことばの言語資源が整備され，またコーパスに基づいた言語研究の手法が開発・確立されていくことによって，話しことばの定量的研究は再び活発になっていくものと思われる．またそのような研究は，話しことばを対象とする自然言語処理技術の研究・開発にとっても，資するところが大きいと考えられる．

コーパスを利用して話しことばを研究する手法の有利な点は，母語話者の内省に基づいて分析を行なう従来の文法研究では気づかれてこなかった (あるいは看過されてきた) さまざまな言語現象を拾い上げ，定量的に分析できるところにある．内省に基づいて言語能力を記述するのではなく，現実世界における言語行動を直接の観察対象とすることによって，例えば，発話の形式と発話者の心的状態・発話の情報構造との相関を明らかにし，言語行動のパタンとして記述するという研究が期待できる．そのような方法論は，外的世界に具現化した外在言語，および発話者の心的状態の両者を自然現象として捉え，それらを有機的に関連付けていこうとする点において，「自然科学的言語研究」とでも呼ぶべきものである．そのケーススタディとして，本稿では話しことばに現れる「挿入構造」という現象に注目し，CSJ に出現した例をもとに分析を行なう．

2　問題の所在

2.1　現象

話しことばは，発話という一回的な行動の中で実時間的に組み立てられていく言語様式である点で，書きことばとは異なる．発話の産出は時間軸に沿って線条的に行なわれるため，何らかの理由によって発話の形式が途中でくずれたり，間違いに気づいて発話し直したりして，当初想定されていた発話の形式や内容が変容を受けることがある．話しことばに特有に見られる言い差し・言い直し・倒置・発話のねじれなどの現象はその例であり，本稿で取り上げる「挿入構造」も，発話の動的な生成という特性によって生じる現象の一つとして位置づけられる．

挿入構造とは，典型的には，ある発話の途中に別の発話が入り込んだ構造のことを指す．発話を行なっている途中に別の発話が挿入された結果，発話開始時から挿入された部分の末尾までを一まとまりと考えると文法的・意味的な不整合が生じるものである．CSJ に出現した挿入構造の例を，以下に挙げる．なお，（F ）で囲まれた部分はフィラー要素を，<p> は 200 ミリ秒以上のポーズが挿入されている箇所を，それぞれ表す．

(1)　私が大学の二年の時に （F えーっと） 千九百 <p> 九十六年の七月です <p> （F えー） <p> 学科の仲間と一緒にキャンプに <p> 行ったことについて <p> 話します

(2)　最後に （F あのー） （F ま） 会長からですね <p> （F あの） （F ま） やめる時の話なんですが 言われたことは <p> お前悔い残すぞと

(1) には，途中に文末表現「七月です」が現れている．しかし，発話開始時から最初に現れる文末表現まで，つまり「私が大学の二年の時に千九百九十六年の七月です」を一まとまりと見ることは，文法的にも意味的にも適切ではない．むしろ，「私が大学の二年の時に学科の仲間と一緒にキャンプに行ったことについて話します」という一つの発話の途中に「千九百九十六年の七月です」という別の発話が入り込んだものと考える方が妥当である．

また (2) には，途中に「やめる時の話なんですが」という従属節が現れている．南不二男 [4] によれば，「が」で

導かれる従属節は従属度の低い従属節であり，その前に現れるニ格やカラ格は「が」節の内部に収まるはずである．しかし，「最後に会長からやめる時の話なんですが」を一まとまりと見るのは解釈として妥当でない．むしろ，「最後に会長から言われたことは」という発話の途中に「やめる時の話なんですが」という別の発話が挿入されたと考える方が自然である．

いずれの例も，発話者がある発話を開始したものの，発話内容に関して補足を加える必要があると考えたため，その場で別の発話 (下線部) を元の発話の途中に挿入して情報を補足し，その後は再び元の発話を再開する，という格好になっている．端的に言えば，「話が途中で脱線して，また元に戻る」ということである．

2.2 先行研究

話しことばに現れる挿入構造という現象について，従来の文法研究はほとんど注意を払ってこなかった．その中で，話しことばの定量的な収集・分析を行なった先述の国立国語研究所 [2] には，次のような記述がある．

> 統一体としての文という基準に照らせば，挿入を持つ文は，ともかくも，くずれたものといわなければならない．しかし，くずれをもちながらそれを包んで統一性を得ているもの，くずれを包んだ統一体というべきであろう．(p.29)

「話しことばに多く現れる (同 p.29)」挿入構造は，文法的には破格であるものの，情報伝達上の機能としては統一的なまとまりを備えているというわけである．

また，(2) のように「が」で導かれる従属節が挿入される場合について，三上章 [5] には次のような記述がある．

> (挿入された)「ガ」の受ける語句は，前置きであり断り書であって，主文の内容とは無関係である．無関係だから順接でも逆接でもなく，ただの平接である．シンタクス上は間投的なユウ式である．上下に切離して，下を「ガ」や「シカシ」で始めるわけにはいかない．(p.302)

ユウ式とは従属節が「間投的に使われる場合の特称」で，「やや遊離した位置にあり，主文への係り方が自由で，意味上誘導の役割をする (同 pp.189–190)」ものである[1]．「が」の主節に対する係り方 (機能) によってその統語的な位置が決まるという見方は実に卓見であり，三上の観察力の鋭さを示していると言ってよい．しかし，このような見方は，その後の研究 (南不二男 [4] など) にはほとんど引き継がれなかった[2]．

話しことばに現れる挿入構造という現象について，以上に見たような断片的な記述はあるものの，実際の話しことばから挿入構造を収集して定量的に分析を行なった研究は，管見の限り見られない．そこで本稿では，CSJ から挿入構造の具体的な事例を収集し，形式ごとの出現数や出現傾向，談話内で果たす機能などについて検討を行なう．

3 分析

3.1 挿入構造の認定

CSJ には句点が書き起こされていないため，係り受け解析，談話構造分析，自動要約などを行なうには，それらに用いられる基本的な処理単位をあらかじめ抽出しておく必要がある．我々は，発話中に現れる「節」の終端境界を手がかりとして自動的に発話分割処理を行ない，その結果を人手で修正することによって，処理単位を抽出・特定する作業を行なっている [7]．この作業には，挿入構造を人手で認定し，挿入された部分の始点と終点をマークするという作業が含まれている．以下の分析では，この作業によって認定された結果を用いる．

人手による挿入構造の認定基準は，以下の 4 点である．

1. その前後に係り受け関係を持つ要素があること．
2. その終端で発話を区切ると，その前後の要素間の文法的・意味的な整合性が保てないこと．
3. 元の発話に対する「前置き」や「断り書き」になっており，その部分を除いても影響がないこと．
4. 挿入された部分の末尾の形式が，文末表現，または接続助詞「けれども[3]，が」であること．

これらの基準について，次の例を用いて検討しよう．

(3) ホテルの <p> 部屋の中も早速 (F あの) 夜着いたんですけども チェックしました

(4) 日曜日になるとマーケットに売りに行ったりしていました <p> 後ホームステイもしたんですけれども <p> ホームステイは <p> 一家族二人 <p> 日本人は二人ずつお世話になりまして <p>

(5) 今日お話しするのは ここに <p> タイトルにありますように (F えっと) 不特定話者 <p> の <p> 認識です

(3) に示した下線部は，前後に係り受け関係を持つ要素 (「中も」「早速」と「チェックしました」) があり，その終端で発話を区切ると文法的・意味的に整合的でなくなる．さらに，元の発話の前置きとなる情報を提示しているため，下線部を除いても元の発話の大意には影響がない．以上から，(3) の下線部は挿入構造と認定される．

一方 (4) では，下線部の前後に係り受け関係を持つ要素はなく[4]，その終端で発話を区切っても文法的・意味的な

[1] ただし，三上は「が」の主節に対する係り方を重視しているため，次のように文頭に現れる「が」の従属節もユウ式に含めている．

チョツトオ伺ヒシマスガ，郵便局ハドチラデセウカ?

本稿で扱う挿入構造は，後述するように，文中に現れる場合のみを対象としており，この点で三上の記述とは異なる．

[2] 従属節の機能と統語的な位置の連動については，丸山 [6] も参照．

[3] 「けれど，けども，けど」という異形態を含む．

[4] 下線部の直前は文末表現であり，発話分割位置となるため，「けれども」節は「文頭」に現れたものであることになる．

不整合は生じない．このため，元の発話の前置きを提示する点では (3) と等しいものの，挿入構造とは見なさない．

さらに (5) では，下線部の前後に係り受け関係を持つ要素があり，元の発話に対する断り書きになっているものの，本稿で扱う形式としての基準を満たさないために，挿入構造とは見なさない．

3.2 結果

CSJ に含まれる 188 講演 (学会講演 77 講演，模擬講演 111 講演) を対象に，人手で挿入構造の認定を行なった．認定作業の結果，合計 821 例の挿入構造が認定された．認定された挿入構造を，挿入部分末尾の形式ごとに分類し，それぞれの出現数と割合，学会講演と模擬講演の内訳を求めた．結果を表 1 に示す．両講演の間で，出現数の順位に若干の違いがある点に注意されたい．

表 1: 認定された挿入構造の内訳

末尾形式		出現数		(学会 / 模擬)
けれども		389	(47.4%)	(104 / 285)
が		214	(26.1%)	(125 / 89)
文末表現	フィラー文	135	(16.4%)	(20 / 115)
	終助詞	71	(8.8%)	(36 / 35)
	その他	12	(1.5%)	(7 / 5)
合計		821	(100%)	(292 / 529)

出現数については，「けれども」が半数近くを占め，「が」と文末表現はほぼ同数であった．なお，「フィラー文」とは，「そうですね」「すみません」「何て言うんでしょう」など，形式上は文末表現になっているが全体でフィラーとしての機能を果たす定型表現のことである．

(6) 横浜まで行くのに そうですね 大体車で一時間半ですね

このようなフィラー文は，学会講演に比べて模擬講演に多く現れている．特に，次に何を喋るかを逡巡しながら発話を続けるような発話スタイルを取る場合に，頻出していると考えられる．

次に，「けれども」「が」の全出現数と，挿入構造として用いられる場合の割合を調べた．結果を表 2 に示す．

表 2: 「けれども，が」の全出現数と挿入構造になる割合

	全出現数 (学会 / 模擬)	挿入構造になる割合 (学会 / 模擬)
けれども	2,782 (680 / 2,102)	13.98% (15.29% / 13.56%)
が	1,948 (1,206 / 742)	10.99% (10.36% / 11.99%)

「けれども」「が」の全出現数は，学会講演と模擬講演の間で大きく異なっていた．しかし，ここで注目したいのは，「けれども」と「が」が挿入構造として用いられる割合が，両講演とも，そして「けれども」「が」ともに，10% ～ 15% の間に収まっており，顕著な差は見られないという点である．つまり，「けれども」「が」が挿入構造として用いられる割合は，講演の種類に関わらず，ほぼ一定していると言うことができる．

さらに，特徴的に観察されたのは，挿入構造の「けれども」「が」の直前に「～んです」のようなノダ形が現れる割合が多いという点であった．「けれども」「が」にノダ形が前接する割合について，表 3 に示す．

表 3: 「けれども」「が」にノダ形が前接する割合

	学会	模擬
けれども	44 / 104 (42.31%)	196 / 285 (68.77%)
が	45 / 125 (36.00%)	61 / 89 (68.54%)

特に模擬講演においては，ノダ形が前接する割合が「けれども」「が」ともに 70% 近くを占めている．「説明のモダリティ [8]」を表すノダ形が挿入構造の述語になりやすいというこの結果は，挿入構造が発話の内容を補足・説明するために機能している場合が多いことの傍証として考えることができる．

4 考察

先述した通り，挿入構造とは，ある発話の途中に入り込み，元の発話に対する「前置き」や「断り書き」として機能するものである．しかし，その挿入構造が談話内でどのような機能を果たしているかという点に着目すると，これらはさらに細かく分類できると思われる．そこで，認定された挿入構造のうち「フィラー文」を除く 686 例について分析したところ，大きく 3 つの下位類型を認めることができた．以下では，3 つの類型 (「A 型」「B 型」「C 型」と呼ぶ) の出現数，および談話内で果たす機能について検討を行ない，挿入構造の類型と発話者の心的状態，および発話全体の情報構造の関係について述べる．

A 型： 発話全体の背景や前提となる情報を表す

(7) 色んなパターンを <p> ここに書いてある数字は頻度ですが たくさん集めてみました

(8) 基本的には （F ま） （F えー） 遠隔教育サーバーというものが （F ま） 一番上の方ですけれども ございまして <p> （F え） そこには （F まー） 教材コンテンツであるとか

B 型： 直前に発話した語や内容について注釈を加える

(9) 正直言って <p> 学部 <p> 私工学部だったんですけど そちらの勉強は <p> 殆ど <p> （F えーと） <p> しておりませんで

(10) 一つは （F え） 全体四つから五つの班 パーティーと言いますが <p> パーティーに分けられて <p> 各パーティーごとに行動を取ります

C 型： 今から発話しようとする語や内容についてあらかじめ注釈を加える

(11) お酒と <p> メニューは少ないんですが 食事が置いてあります

(12) 状態述語というのは <p> 例文の十二を見てください <p> 十二番のように （F えー） 述語が形容詞のものと十三番のように動詞のものがあります

各類型の出現数を，形式ごとに表 4 に示す．

表 4: 各類型の出現数と形式

	出現数	(学会 / 模擬)	けれども	が	文末
A 型	268	(114/154)	149	97	22
B 型	238	(97/141)	131	69	38
C 型	180	(61/119)	109	48	24
合計	686	(272/414)	389	214	83

A 型と B 型における発話者の心的状態は，当該の発話を開始した後に発話の前提となる情報や直前に発話した内容に関する情報について注釈を加える必要があると気づいたため，その場で急遽発話を中断して補足情報を強制的に挿入し，その後また元の発話を再開する，というものである．特に B 型では，挿入が終わって元の発話に戻る際，(9) の「そちらの」，(10) の「パーティー」のように，挿入直前に述べた要素を指示したり，言い換えのために挿入した語を繰り返したりすることが多い．

B 型にはさらに，所定の位置に別の発話を挿入することがあらかじめ予定されているように思われる場合もある．

(13) まずさまざまな空間音響特性を学習したエイチエムエム <p> （F え） これを空間音響特性依存エイチエムエムと呼びます <p> を複数用意します

下線部の挿入部は，発話中に急遽挿入されたものでなく，「... 学習したエイチエムエム」の直後に挿入することが発話開始時にあらかじめ想定されていたと思われる．挿入が事前に準備されていた点で，発話がその場で中断される典型的な挿入構造とは異なる．このような挿入構造は学会講演によく見られ，特に (13) のように何かを定義するための文や，(14) のように発話した内容を言い換えたり解説したりするための名詞述語文が多く見られた．

(14) スコアが高ければ高い程 <p> （F え） 原型と変形の聴取時間の差 <p> これは差の絶対値 <p> です <p> が （F えー） 高水準で相関が認められました

C 型は，これから発話しようとする内容について前もって補足情報を提示するという点で，A，B 型と異なる．発話はすでに開始しているものの，情報を補足すべき要素にはまだ言及していないため，これから発話する内容の前置きとして元の発話に挿入される格好になっている．補足する要素がまだ現れていないことから，B 型のように挿入部の直後に指示要素が現れることは少ない．

以上のように，発話内における挿入の位置の違い，および談話内で果たす機能の違いは，発話者の心的状態や発話の情報構造の違いを反映する．一旦開始した発話の途中に別の発話を割り込ませることにより，発話全体の情報構造が柔軟に操作されるわけである．発話は基本的に一次元的・線条的な産出物であるが，発話者は挿入構造を用いることによって情報の立体的な提示を実現し，聞き手に対する情報伝達の効率化を図っていると考えられる．

なお，文末表現が現れた後，その直前に現れた名詞を項にする助詞が現れ，さらに発話が続く場合がある．挿入構造とは異なるが，関連する現象として指摘しておく．

(15) 父が <p> （F えーと） <p> （F ま） 製鉄所ですか <p> に （F あのー） <p> 勤めていた関係で

(16) 友人 <p> のエム君 <p> が 交通事故 <p> になるんでしょうね <p> で死にました

5 結語

話しことばに現れる現象を定量的に収集して分析を行なう研究手法のケーススタディとして，挿入構造という言語行動の様式について，CSJ に現れた実例をもとに分析を行なった．本稿では，挿入構造の形式を文末表現，接続助詞「けれども」「が」のみに限定したが，他の従属節や句が挿入される場合も考えられるため，より多様な形態について検討することが必要である．また，挿入構造が談話内で果たす機能についても，言い直しや言い差し，倒置など関連する現象も含めて網羅的に分析することにより，より詳細な機能的類型を設けることができると思われる．さらに，挿入された部分の音韻的特徴や，挿入部の直近に現れるフィラー・ポーズの分布などを射程に入れることにより，音声的側面から挿入構造を分析することが可能になると思われる．いずれも今後の課題としておく．

謝辞: 本研究は，通信・放送機構の研究委託により実施したものである．

参考文献

[1] 国立国語研究所： “話しことばの文型 (1)”, 秀英出版 (1961).

[2] 国立国語研究所： “話しことばの文型 (2)”, 秀英出版 (1963).

[3] 前川： “『日本語話し言葉コーパス』の設計と実装”, 平成 15 年度国立国語研究所公開研究発表会予稿集 (2003).

[4] 南： “現代日本語の構造”, 大修館書店 (1974).

[5] 三上： “現代語法序説”, 刀江出版 (1953).

[6] 丸山, 熊野, 柏岡： “日本語における独話の特徴と文分割”, 言語処理学会第 7 回年次大会 発表論文集, pp. 429–432 (2001).

[7] 高梨, 丸山, 内元, 井佐原： “話し言葉の文境界 –csj コーパスにおける文境界の定義と半自動認定 –”, 言語処理学会 第 9 回年次大会発表論文集, pp. 521–524 (2003).

[8] 益岡： “モダリティの文法”, くろしお出版 (1991).