

自発的な話し言葉に現れる言い直し表現の機能的分析

丸山 岳彦[†]

[†] 独立行政法人 国立国語研究所

佐野 真一郎[‡]

[‡] 上智大学 大学院 外国語学研究科

1 導入

音声認識や音声翻訳、音声要約など実際の話し言葉を処理しなければならない音声言語処理研究においては、実質的な要請から、話し言葉に頻出する非流暢現象への対処が研究されてきた。例えば、自発的な話し言葉には「言い直し表現」が頻繁に観察されるが、その検出方法と処理について、いくつかの先行研究がある [5, 6]。一方、言語学的な研究においては、非流暢現象は言語運用上のノイズとして見なされ、研究対象として取り上げられることはほぼ皆無であった。

2004 年の『日本語話し言葉コーパス』の完成・公開 [4] により、実際の話し言葉を定量的に収集・観察し、従来ノイズと見なされてきた現象を言語学的な視点から分析するための環境が整ったと言える。そこで本稿では、自発的な話し言葉に現れる言い直し表現を取り上げ、それらが談話内でどのような機能を担うかという問題について考察する。以下では、『日本語話し言葉コーパス』から言い直し表現を手作業で収集し、その談話的な機能を 5 つの類型に分類した結果について示す。さらに、その数量的な分布について分析を行う。

2 言い直し表現とは何か

自発的な話し言葉において、話し手は自らの発話や聞き手からの反応を常にモニタリングしている [1]。自身の発話内容に何らかのトラブルを検知した場合、話し手はできるだけ速やかに発話内容を修復しなければならない。発話中に現れたトラブルを修復するためにある一定の形で発話される表現が、本稿で扱う言い直し表現である。言い直し表現の例を (1) に挙げる。「(F)」「(D)」で囲まれた要素は、CSJ で記述されたフィラー、語断片にそれぞれ相当する。

- (1) a. 異なる (D アク) (F あ) アクセント型の 語
b. (F えー) 黒い (F あー) 失礼 青い 丸は
c. クレーターが とても 大きな クレーターが ありまして

(1a) は、「アクセント型」と言うべきところを「アク」と言い淀んでしまい、改めて「アクセント型」と言い直す例である。(1b) は、「黒い」が間違いであることに気づき、「青い」と言い直す例である。(1c) は、「クレーターが」と言った後、「とても大きな」という修飾表現を補足して全体を言い直す例である。いずれも、一度発話した内容を言い直す点では共通している。しかしながら、発話した内容のどの側面が言い直されているかという点においては、いずれの例も異なっている。このことから、言い直し表現には、談話内で担う機能という点において、複数の類型が認められると言える。

話し手は、自らの発話内容にトラブルを検知した場合、文法構造の適格性に破綻を生じさせることと引き換えに、言い直し表現を用いて発話の修復を行わなければならない。発話という行為が一時的な行動の結果である以上、そこに何らかのトラブルが発生するのは不可避的である。この点において、言い直し表現とは、発話に不可避的に現れるトラブルをその場で修復し、より正確かつ適切な情報伝達を行うための言語編集方略としての役割を担うと言える。

3 分析対象 — 『日本語話し言葉コーパス』

本稿では、『日本語話し言葉コーパス (The Corpus of Spontaneous Japanese; 以下 CSJ)』を分析データとして用いる。CSJ のうち、「コア」と呼ばれるセットに含まれる独話 177 講演 (「学会講演」70 講演、「模擬講演」107 講演) を分析対象とする。また、各講演の講演者に対するアンケートを収録した「記録票データ」を用いる。CSJ の詳細については、文献 [4] を参照されたい。

4 言い直し表現の構造と機能的分類

以下では、本稿で提案する言い直し表現の構造と機能的分類について示す。

4.1 言い直し表現の構造化とアノテーション

本稿では、Nakatani and Hirschberg(1993) による“Repair Interval Model(RIM)”[3]などを参考にして、言い直し表現の構造を次のような図式で捉える。

(2) (被言い直し部 | 中断部 | 言い直し部)

被言い直し部は、後に続く言い直し部で修正されたり言い換えられたりする対象である。被言い直し部には、語断片、語、句、節などさまざまな要素が現れ得る。中断部は、被言い直し部と言い直し部の間に随意的に挿入される部分で、フィラー「えーと」「あのー」やメタ的な編集表現「と言うか」「何だっけ」「すみません」などが現れる。言い直し部は、被言い直し部を適切な形で修正したり言い換えたりする部分である。

アノテーション方法として、被言い直し部の冒頭から言い直し部の末尾までを“()”で囲んだ。中断部の前後の境界には、“|”を挿入した。これらを付与する単位は、形態論情報の「短単位」を基準とした。言い直しを認める範囲は、CSJに付与されている「節単位」を超えない範囲とした。また、言い直しが重層的に起こっている場合、入れ子構造を作ること認めた。

以上のアノテーションに基づいて、CSJのコアに含まれる独話177講演を観察し、言い直し表現を収集した。収集作業は、実際の音声聞きながら、転記テキストおよび形態論情報(短単位)を参照して行った。以下、「語」という場合は短単位を指す。

4.2 言い直し表現の機能的分類

次に、言い直し表現を機能的な観点から分類する。収集された言い直し表現の事例を観察し、その言い直し表現が談話中でどのような機能を担っているかという観点から、探索的に検討を行った。その結果、収集した言い直し表現は、次の5つの類型に分類できた。

(3)

| |
|-------------------|
| R1: 発音エラーに伴う言い直し |
| R2: 単純な繰り返し |
| R3: 語彙的な誤りに伴う言い直し |
| R4: 情報不足に伴う言い直し |
| R5: 別表現への言い換え |

これらのタイプの別は、(4)のように、言い直しの冒頭部分にR1~R5のタグを付与することで区別した。

(4) (R1 (D アク) | (F あの) | アクセント)

以下では、それぞれの類型について記述する。

R1: 発音エラーに伴う言い直し

「発音エラーに伴う言い直し(R1)」は、発音上の誤りとその修復に相当する。これは話し手が誤った発音をしてしまい、それを直後に修正する例である。被言い直し部には、常に語断片が存在する。

- (5) a. コンテキスト (R1 (D いぞ) | | 依存) モデルを
b. (R1 私の 親の (D おは) | | 私の 親の お墓) が

R2: 単純な繰り返し

「単純な繰り返し(R2)」は、一度発話した内容をそのまま繰り返す場合に相当する。ここで認める繰り返しは、発話中の逡巡やためらい、自己確認などに起因して生じるものであり、意図的な繰り返しは対象としない。意図的な繰り返しになっているか否かは、音声参照して判断した。

- (6) a. その 年の (R2 七月 | | 七月) から
b. ちょっと (F あ) (F ま) (R2 散らばっている | | と言いますか (F あの) (F あ) (F ま) | | 散らばっている) 訳です

R3: 語彙的な誤りに伴う言い直し

「語彙的な誤りに伴う言い直し(R3)」は、誤った語や表現を発話してしまい、それを直ちに修復する場合に相当する。この場合、誤りをそのまま放置すると聞き手に不正確な情報が伝わってしまうため、話し手は誤りに気づいた時点で直ちに修復を行う必要がある。

- (7) a. 同音異義を (R3 分別 | (F えー) | 弁別) している という 風に
b. (R3 千九百九十九 | (F あ) | 千九百九十) 年の
(8) a. 独特の (R3 旋律の | (F え) | 旋律を) 形作っている
b. 音名 (R3 (D2 の) | (F え) | に) 相当する ものです
(9) a. (R3 こちらの 条件は | (F あ) | (R1 (D こ) | | こちらの) の 被験者は)
b. (R3 原型を (F えー) 有意に 聞いた | | 変形を 有意に 聞いた) という 人数比を

(7a)は、「弁別」と言うべきところを誤って「分別」と発話してしまい、それを修復する例である。(8a)は「旋律を」と言うべきところを誤って「旋律の」と言っ

てしまい、それを修復する例である。この場合は、助詞の選択誤りを言い直しているため、名詞「旋律」の意味役割が言い直されていることになる。これは、(8b)のように、助詞だけが言い直される場合もある。さらに、(9)のように、句や節がまるごと言い直される場合もある。

R4: 情報不足に伴う言い直し

「情報不足に伴う言い直し (R4)」は、一度発話した内容の情報量が不十分であったために、補足の表現を加えた上で言い直す場合に相当する。被言い直し部に本質的な誤りはないものの、補足する情報を加えてより適切な形で言い直す場合である。

- (10) a. 入ってきた 途端に (R4 ショーケース || ドーナツのショーケース) を見て
b. 水疱を 伴った (R4 ぶつぶつ || じんま疹 みたいな ぶつぶつ) が (F え) 突然 できる
- (11) a. 海外に (R4 興味 || 元々 興味) が ありまして
b. (R4 エヌ個に | (F え) | ケーミーンズ法によって エヌ個に) (F えー) 分割します

(10a) は一度「ショーケース」と発話したものの、それを補足する表現が不足していると気づき、「ドーナツの」という連体要素を補足した上で全体を言い直している。(11a) は「興味」と発話したものの、さらに後方の述語を修飾する要素をまず言うておく必要があることに気づき、「元々」という連用要素を補足した上で全体を言い直している。この種の言い直しは、発話の動的な構築過程において、構文要素を線状的に配列する際、その順序が混乱することにより生じるものである。

R5: 別表現への言い換え

「別表現への言い換え (R5)」は、一度発話した内容を別の表現で言い換える場合である。厳密には、発話中に発生したトラブルとその修復という観点からは外れるものであるが、R1~R4 と同じ構造を持つこと、場合によっては適格な文法構造の破綻を引き起こすことなどから、言い直し表現の一つとして考えておく。

- (12) a. (R5 オクターブ 上 に なった | つまり | 周波数が二倍に なった) 時に
b. (R5 長三和音 | 例えば | ドミソミドの (F えー) 分散和音) 型を ずっと 聞かせ続けて
c. この イータが 大きい 程 (R5 収束が || コンバージェンスが) (F ま) 速くなる
d. 左側の グラフは (R5 エルエルエックス | つまり | 下側の 唇の 前後 位置) を 表わし
e. (R5 こちら側の グラフ || 右側の グラフ) は 何を 表わしているかと言いますと

言い換えのパターンには、発話した内容を別の表現で解説する場合 (12a)、発話した内容の具体的な例を言い足す場合 (12b)、専門用語への言い換え (12c) や専門用語からの言い換え (12d)、発話の場において現場指示を行うための言い換え (12e) などがある。

5 分析 — 言い直し表現の分布

以下では、177 講演に現れた言い直し表現の出現数の分布、言い直し表現の長さ、話し手の講演に対する自己評価と言い直し表現との関係について分析する。

5.1 分析 1: 言い直し表現の出現数

まず、言い直し表現の出現数の分布について見る。アノテーションの結果、合計 5,961 例の言い直し表現を収集することができた。また、学会講演・模擬講演の別に、言い直し表現の総数を総語数で割ることにより、どれだけの比率で言い直し表現が出現しているかを示す「言い直し率」を求めた。結果を表 1 に示す。

表 1 を見る限り、学会講演と模擬講演の言い直し率はほぼ等しい分布となっている。言い直し表現の類型ごとに見ても、語彙的な誤りに伴う言い直し (R3) が学会講演で若干高い割合を示すものの、両者の分布はほぼ等しいと言える。

5.2 分析 2: 言い直し表現の長さ

次に、言い直し表現の長さの分布について見る。言い直しが重層的に起こっていない 5,425 例を取り出し、それぞれの被言い直し部・言い直し部にいくつの語が含まれているかを調べた。学会講演・模擬講演の別に、平均語数の分布を図 1、2 に示す。

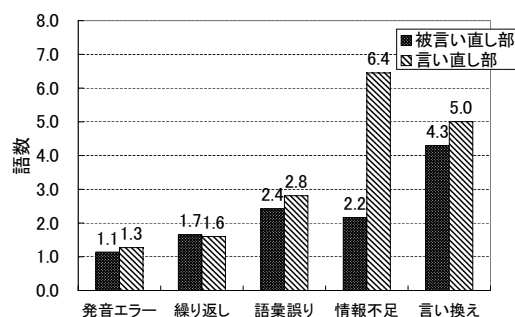


図 1: 言い直し表現の長さの分布 (学会講演)

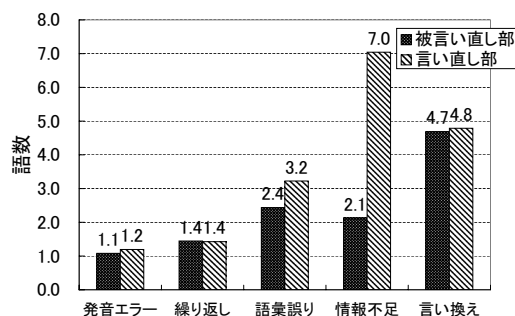


図 2: 言い直し表現の長さの分布 (模擬講演)

表 1: 言い直し表現の出現数と言い直し率

| | 総語数 | R 総数 | R1 | R2 | R3 | R4 | R5 |
|------|---------|-----------------|-----------------|---------------|-----------------|-----------------|-----------------|
| 学会講演 | 220,865 | 2,940 1.331% | 724 0.328% | 228 0.103% | 902 0.408% | 544 0.246% | 542 0.245% |
| 模擬講演 | 226,902 | 3,021 1.331% | 746 0.329% | 263 0.116% | 809 0.357% | 606 0.267% | 597 0.263% |
| 合計 | 447,767 | 5,961 1.331% | 1,470 0.328% | 491 0.110% | 1,711 0.382% | 1,150 0.257% | 1,139 0.254% |

(上段は出現数、下段は言い直し率)

局所的な発音エラー (R1) から大局的な言い換え (R5) へと移るにつれて、言い直し表現の長さが長くなっていることが確認できる。情報不足 (R4) の言い直し部が突出して長いのは、新しい情報を追加した上で全体を言い直すという R4 の特徴によるものである。

5.3 分析 3: 講演の得手・不得手と言い直し率

最後に、話し手が自分がどれだけ講演が得意であるかを評定した結果と、言い直し表現の出現率との関係について見る。CSJ には、講演者に対して事前にアンケートを取った結果が「記録票データ」として記録されており、この中に「講演の得手・不得手」という項目がある。これは自分が講演が得意かどうかを回答させたもので、「得意」「やや得意」「やや不得手」「不得手」という 4 段階で回答されている。この基準により学会講演・模擬講演の別に講演を 4 つのグループに分け、言い直し率の分布を調べた。結果を図 3、4 に示す。

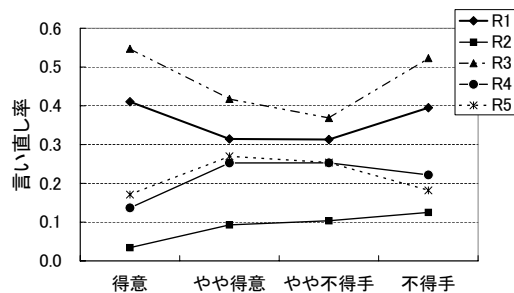


図 3: 講演の得手・不得手と言い直し率 (学会講演)

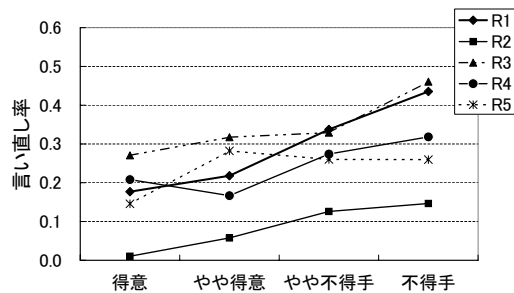


図 4: 講演の得手・不得手と言い直し率 (模擬講演)

学会講演では得手・不得手と言い直し率の関係は見取れないが、模擬講演では話し手が不得手と回答するほど言い直し率が上昇していることが読み取れる。少なくとも模擬講演の場合、実際の言い直し表現の多寡は、得手・不得手の度合いに対する話し手の意識と連動していると見ることができる。

6 結語

本稿では、CSJ の 177 講演に現れた言い直し表現を手で収集した。また、収集された 5,961 の言い直し表現について、その談話的機能を 5 つの類型に分類し、タグ付与を行った。さらに、各言い直し表現の出現数の分布、言い直し表現の長さ、講演者の自己評定と言い直し表現の多寡について分析を行った。

CSJ には、本稿で取り上げた分析項目以外にも、発話スピード、スタイル、自発性、緊張度などさまざまな言語外要因に関するデータが記録されており、これらの要因と言い直し表現の分布の関連についても検討中である。さらに、言い直し表現の機能に関する類型の細分化や、より詳細な統計的分析なども、今後の課題として挙げられる。

付記: 本研究は、科学研究費補助金「現代日本語の話し言葉に現れる「言い直し表現」の体系化に関する研究」(若手研究 (B)、課題番号 17720098) の補助を受けている。

参考文献

- [1] Levelt, W. (1989) *Speaking: From Intention to Articulation*. The MIT Press.
- [2] Maruyama, T and Sano, S. (2006) Annotation and Analysis of Self-Repairs in Japanese Spontaneous Monologues, *Proc. of Linguistic Patterns in Spontaneous Speech*.
- [3] Nakatani, C. and Hirschberg, J. (1993) A Speech-first Model for Repair Identification and Correction, *Proc. of 31th Annual Meeting of ACL*.
- [4] 国立国語研究所 (2006) 『日本語話し言葉コーパスの構築法』。(国立国語研究所報告書 124)
- [5] 伝康晴 (1997) 統一モデルに基づく話し言葉の解析, 『自然言語処理』. 4(1). 言語処理学会
- [6] 船越孝太郎他 (2003) 音声対話システムにおける日本語自己修復の処理, 『自然言語処理』. 10(4). 言語処理学会