# 日本語の言い直し表現に対するアノテーション基準

吉田　奈央[1]　丸山　岳彦[1]
[1] 専修大学
nao.yoshida@senshu-u.ac.jp　maruyama@isc.senshu-u.ac.jp

## 概要

現代日本語の話し言葉に現れる「言い直し表現」（self-repair）について、その種類を認定してアノテーション（情報付与）する基準と方法論を議論する。国語研 [1] による『日本語話し言葉コーパス』（CSJ）の独話音声（約 44 時間分）を観察し、そこに現れる言い直し表現の認定・アノテーションを実施した。同コーパスから言い直し表現を収集・分類した先行研究 [2] では拾いきれなかった事例や、その細分類・再分類も含めて検討し、言い直し表現の認定基準について議論を行った。

## 1　はじめに

日常の話し言葉の中では、流暢な発話産出の過程がさまざまな要因により阻害される（トラブルが起こる）ことがある。流暢な発話産出にトラブルが生じたことを検知した話し手は、発話産出の流暢さを取り戻すべく、できる限り速やかにトラブルの修復に取り掛かる。前者の「流暢な発話産出の阻害要因」には、言い誤り、発音エラー、語の選択誤りなどが含まれ、後者の「発生したトラブルの修復」には、繰り返し、言い直し、置き換え、追加、挿入などが含まれる。これらは、発話産出過程で不可避的に生じるトラブルと、流暢な発話産出を継続するための方略、という図式で捉えることができる。これらをまとめて、非流暢性（disfluency）と呼ぶ。

本稿では、非流暢性のうち「言い直し」（self-repair）を取り上げ、丸山（2008）[2] の分類案を一部改訂する形で、談話中における言い直しの形態・機能を 5 種類に分類する。ここでは、丸山（2008）[2] にならい、言い直しを「発話中に現れたトラブル、およびそれを修復するために発話される表現」と定義し、国語研 [1] による『日本語話し言葉コーパス』（CSJ）の独話音声（44 時間分）を観察し、そこに現れる言い直し表現の認定・アノテーションを実施した。その過程で検討した言い直しの認定基準や、丸山（2008）[2] の細分類・再分類について、議論を行う。さらに、当該の言い直しが偶発的に生じたものなのか、または発話者によって聞き手への配慮として意図的に行われた言い直しなのかについて、これらは動機が異なる点で別の事例として捉える必要があることを論じる。

## 2　先行研究

### 2.1　言い直しをどう捉えるか

1970 年代の社会学における会話分析や、言語心理学、会話ストラテジーの研究において、言い直しは、言い誤りを起こした話し手の修復行為として、または聞き手側の承認を得るための会話ストラテジーとして、議論の対象となってきた（Fromkin 1973[3], Levelt 1989[4], Sacks et al. 1974[5], Williams 1997[6]）。一方、1990 年代以降の音声言語処理の研究の中では、言い直しは適切な係り受け解析に対する障害となり得ることから、発話の整形処理の対象として論じられてきた（中野 1998[7], 藤井 2004[8], 下岡 2005[9], 船越 2006[10], 尾嶋 2006[11]）。さらに 2000 年代、日本語の話し言葉コーパスの整備が進んだことから、記述的な言語研究においても非流暢性を定量的に扱う研究が進められている。

### 2.2　非流暢性としての言い直し

その中で丸山（2008）[2] は、日本語の自発的な独話における言い直しを取り上げ、言い直しの構造化、および定量的な分析を行っている。丸山（2008）[2] は、Nakatani and Hirschberg（1993）[12] の RIM モデルなどの先行研究を参照し、言い直しを次の 3 つの区間によって構造化した。「被言い直し部」は言い直しの対象となる区間、「中断部」はフィラーや「ていうか」「失礼」などの編集表現、「言い直し部」は言い直した区間、にそれぞれ該当する。

被言い直し部 | 中断部 | 言い直し部

さらに丸山（2008）[2] は、言い直しの形態的特

徴および談話中で果たす機能という側面から、言い直しを 5 種類に分類した。この分類基準をもとに、『日本語話し言葉コーパス』（CSJ）に対してアノテーションを行い、言い直しの定量的な分析を実施している。

今回、丸山（2008）[2] によるアノテーションを再検討し、その分類基準をさらに明確化した上で、必要に応じてアノテーションされたタグの修正作業を実施した。以下では、その内容について示す。

# 3　言い直しの分類

言い直しの形態的特徴および談話中で果たす機能という側面から、言い直しを以下に挙げる 5 種類に分類する。なお、言い直しを構成する各区間の名称を次のように改め、その範囲を （） で囲むことにする。また、言い直しの分類名を冒頭に記す。

（R1 修復対象区間 | 編集区間 | 修復区間 ）

## 3.1　R1: 発語の失敗に伴う繰り返し

「R1：発語の失敗に伴う繰り返し」は、意図した語を発話（発語）しようとしたものの、その発音に失敗してしまい、それを直ちに（反射的に）修復するケースに相当する。修復対象区間には、語未満の音声（語断片）が現れることが多い。以下の c. では、言い直しが再帰的に生じている。

a. タイプ分け (R1 (D しみゃし) | | しまし ) た
b. 明治神宮 あるいは (R1 新宿 (D くい) | (F あ) | 新宿御苑 )
c. 映画に (R1 (R1 (D なれしんだ) | (F う) | (D なれしたん) ) | | 馴れ親しんだ )

## 3.2　R2：単純な繰り返し

「R2：単純な繰り返し」は、一度発話した語句をそのまま繰り返す場合に相当する。ここで扱うのは意図的な繰り返しではなく、非流暢性としての繰り返しであり、典型的には、発話中の逡巡やためらい、言いよどみ、自己確認に起因するものである。以下の b. c. は、一度発話しかけた語句が中断してしまい、それを改めて繰り返す場合に相当する。

a. (F あのー) (R2 ハワイ | | ハワイ ) って いう の は
b. (F あの) (R2 (D う) | | 運転 ) 免許 が 四 ドル で
c. (R2 被験 (D し) | (F あー) (F あの) | 被験者 ) が

## 3.3　R3：語句の選択誤りに伴う訂正

「R3：語句の選択誤りに伴う訂正」は、話し手の意図とは異なる語句や表現を発話してしまい、それを直後に訂正する場合に相当する。R3 の事例には、R1 や R2 と同様に偶発的に生じたトラブル（語の選択誤り）の場合もあれば、そこまで進んでいた発話を中断し、その場でその後の発話計画を変更する場合や、機能語の訂正を行うことで後続発話に対する係り方を変える場合など、多様なパターンがある。ここでは、後続発話へのつながりを明示的にするため、訂正される対象が内容語か機能語かに応じて「R3-C（内容語）」「R3-F（機能語）」という下位の類型を設けておく。

a. 同音異義を (R3-C 分別 | (F えー) | 弁別) している という
b. 短調独特の (R3-F 旋律の | (F え) | 旋律を) 形作っている
c. クロマというのは 音名 (R3-F (D2 の) | (F え) | に) 相当するものです

## 3.4　R4：不適切な発話に伴う追加と繰り返し

「R4：不適切な発話に伴う追加と繰り返し」は、一度発話した語句では情報として不十分であると気づき、情報を補足する表現を追加した上で、同じ語句を繰り返すケースである。修復対象区間の中に本質的な誤りはないものの、そのままでは聞き手にとって分かりにくいと気づき、十全な発話としては不適切であると判断したものである。

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. (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

## 3.5　R5：不適切な発話に伴う言い換え

「R5：不適切な発話に伴う言い換え」は、一度発話した語句を別の表現を使って言い換えるケースである。この種の修復が起こる動機には、一度発話した語句が聞き手にとって理解が難しいと気づいたために別の平易な表現に言い換える場合や、発話した語句をより具体的な内容を示すために言い換える場合、ある語句を現場指示・文脈指示の表現に言い換えたりする場合などがある。この点で、R4 と同様、聞き手に対する十全な情報提供の適切さに対してト

ラブルを検知し、それを意図的に修復する行為であると言える。

a. (R5 ミクスチャーの | (F あの) | 混合重みの) 係数なんですけれども
b. 正答率は かなり (F ま) (R5 内側に 来ている | | 低くなっている) ことが 分かります
c. 強さパターンの効果は (R5 三．六 | | こちら) (F あ) で 有意な効果が 見られました

# 4 アノテーション

## 4.1 対象データ及びアノテーション手法

今回、『日本語話し言葉コーパス』(CSJ) の独話・コアデータに対し、丸山 (2008) [2] で付与された言い直しのタグをチェックし、必要に応じて修正を行った。言い直しのチェック、タグの修正作業は、日本語文法の専門知識を持つアノテーター 3 名によって実施した。アノテーションにはアノテーターの作業負担を軽減し、作業にかかる誤差をより少なくできるという利便性から、オープンソースライセンスで提供されているアノテーションツール brat[1)] を使用した。丸山 (2008) [2] で作成された言い直しのタグ付き転記テキスト (短単位で区切り、節単位ごとに 1 行にしたもの) を音声を聞きながら確認し、言い直しの範囲およびその種類 (R1〜R5) を同定・確認して、必要に応じて修正を行った。アノテーター間で判断が分かれた場合、合議を行い、合意を得た。3 名の意見が合わない場合は多数決で決定した。これらの作業を通じて得られた、タグを付与するための判断過程を、フローチャートの形にまとめた。Appendix に図 1 として示す。

## 4.2 アノテーション結果

上記の作業の結果、学会講演 70 講演、模擬講演 107 講演 (合計 441,759 語) の中から、5,279 箇所の言い直しを再認定した。一覧を以下に示す。

| R1 | R2 | R3-C | R3-F | R4 | R5 |
|---|---|---|---|---|---|
| 653 | 1449 | 232 | 757 | 1027 | 1161 |

1) https://brat.nlplab.org/

# 5 問題点

## 5.1 言い直しの細分類・再分類

今回のアノテーションの修正作業を通じ、明らかになった問題点は以下の通りである。

### 5.1.1 R3 の動機の分類が不明瞭

今回、「R3：語句の選択誤りに伴う訂正」について、内容語の訂正 (R3-C) と機能語の修正 (R3-F) の 2 つを細分類した。これは、「語句の選択誤り」という点では共通しているものの、それがどのような原因に基づくトラブルであるのか、という点において異なる。先の例を一部再掲すると、

a. 同音異義を (R3-C 分別 | (F えー) | 弁別) して
b. 独特の (R3-F 旋律の | (F え) | 旋律を) 形作って
c. 音名 (R3-F (D2 の) | (F え) | に) 相当するものです

a. は、「分別」と「弁別」を取り違えている例である。これは、発話プランニングの中で語彙選択をする際、よく似た音および意味を持つ語を誤って選択してしまった場合と考えられる。一方、b. は、「旋律の」まで発話した後、その後に続くはずだった統語構造がうまく組み立てられないことに気づき、「旋律を」と助詞を言い直した上で、「形作る」につなげる、という動機が働いているものと考えられる。c. も同様、「の」を「に」に言い換えて、当初に想定していた統語構造に沿って、発話を継続しているものと考えられる。すなわち、同じ「R3: 語句の選択誤りに伴う訂正」の下位分類として設けた R3-C と R3-F は、トラブルの原因としてはかなり異質なものである可能性がある。

### 5.1.2 R5 の動機の分類が不明瞭

これと同様に、「R5：不適切な発話に伴う言い換え」についても、どのような動機によって言い換えが生じているのか、という点で、細分類・再分類が可能になると考えられる。一部を再掲すると、

a. (R5 ミクスチャーの | (F あの) | 混合重みの) 係数
b. (R5 内側に 来ている | | 低くなっている) ことが
c. 効果は (R5 三．六 | | こちら) (F あ) で 有意な

a. は専門用語を平易な表現に言い換える例、b. は観察された現象に対する解釈に言い換える例、c. は言及した数値をスライド上で指し示す例 (現場指

示）である。いずれも聞き手に配慮することによる言い換えであり、談話構成上のストラテジーと考えられるが、どのような配慮のために言い換えを行うのか、という点に着目すると、それぞれは異なる事例と捉えることができるだろう。

動的に進行する発話産出処理において、どの過程がどのような要因でトラブルを起こしているのか、という点に着目すると、言い直しの種類をさらに細分類・再分類することが可能になると考えられる。今回のアノテーション結果に対して、心理言語学的な観点から再分析を行うことによって、さらに多様な言い直しの実態を把握できるだろう。この点に対する分析と考察は今後の課題としたい。

## 6 まとめ

本稿では、現代日本語の話し言葉に現れる言い直しをアノテーションする基準と方法論について議論を行った。丸山（2008）[2] におけるアノテーション結果に対して再検討を行い、新たにアノテーションをし直した。この結果は、一般に公開する予定である。このアノテーションデータをもとに、言い直しの研究がさらに進展することを願いたい。

また、言い直しがどのような動機によって発生するのか、という心理言語学的な観点を取り入れることにより、さらに言い直しのタイプを細分類・再分類することできる可能性を指摘した。この点の検討については、今後の課題としたい。

さらに、今回は『日本語話し言葉コーパス』（CSJ）という独話のデータを分析対象としたが、独話ではなく会話を分析対象とした時にどのような現象が観察されるのか、という点を考えることも重要である。会話中でのインタラクションを考えた場合、発話者の態度や聞き手の承認を得るための、共同作業のストラテジーとして、言い直しが語用論的な役割を担うことがあると考えられる。そもそも今回のアノテーションの方式が会話にも適用できるのか、という点から考え始めることが必要だろう。この点についても今後の課題とする。

## 謝辞

本研究は JSPS 科研費 20H05630 の助成を受けたものです。

## 参考文献

[1] 国立国語研究所. 日本語話し言葉コーパスの構築法. 国立国語研究所, 2006.

[2] 丸山岳彦. 『日本語話し言葉コーパス』に基づく言い直し表現の機能的分析. 日本語文法, Vol. 8, No. 2, pp. 121–139, 1999.

[3] A Fromkin, Victoria. The non-anomalous nature of anomalous utterances. **Language**, Vol. 47, pp. 27–52, 1971.

[4] W. J. M. Levelt. **Speaking From intention to articulation**. MIT Press, 1989.

[5] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. **Language**, Vol. 50, No. 4, pp. 696–735, 1974.

[6] R. Inscoe Williams, J. and T. Tasker. Communication strategies in an interactional context: The mutual achievement of comprehension. **Communication Strategies: Psycholinguistic and Sociolinguistic Perspectives**, pp. 304–322, 1997.

[7] 中野幹生, 島津明. 言い直しを含む発話の解析. 情報処理学会論文誌, Vol. 39, No. 6, pp. 935–1943, 1998.

[8] 藤井なつ音. 日本語話し言葉における自己修復の統計モデル. 第 10 回言語処理学会年次大会発表論文集, 2004.

[9] 下岡和也, 内元清貴, 河原達也, 井佐原均. 日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化. 自然言語処理, Vol. 12, No. 3, pp. 3–17, 2005.

[10] 船越孝太郎, 徳永健伸, 田中穂積. 音声対話システムにおける日本語自己修復の処理. Vol. 10, No. 4, pp. 33–53, 2003.

[11] 尾嶋憲治・河原達也・秋田裕也. 話し言葉の整形作業における削除箇所の自動同定. 情報処理学会研究報告, Vol. 46, pp. 85–91, 2008.

[12] C. Nakatani and J. Hirschberg. A speech-first model for repair identification and correction. **Proc. of 31th Annual Meeting of ACL**, pp. 200–207, 1993.

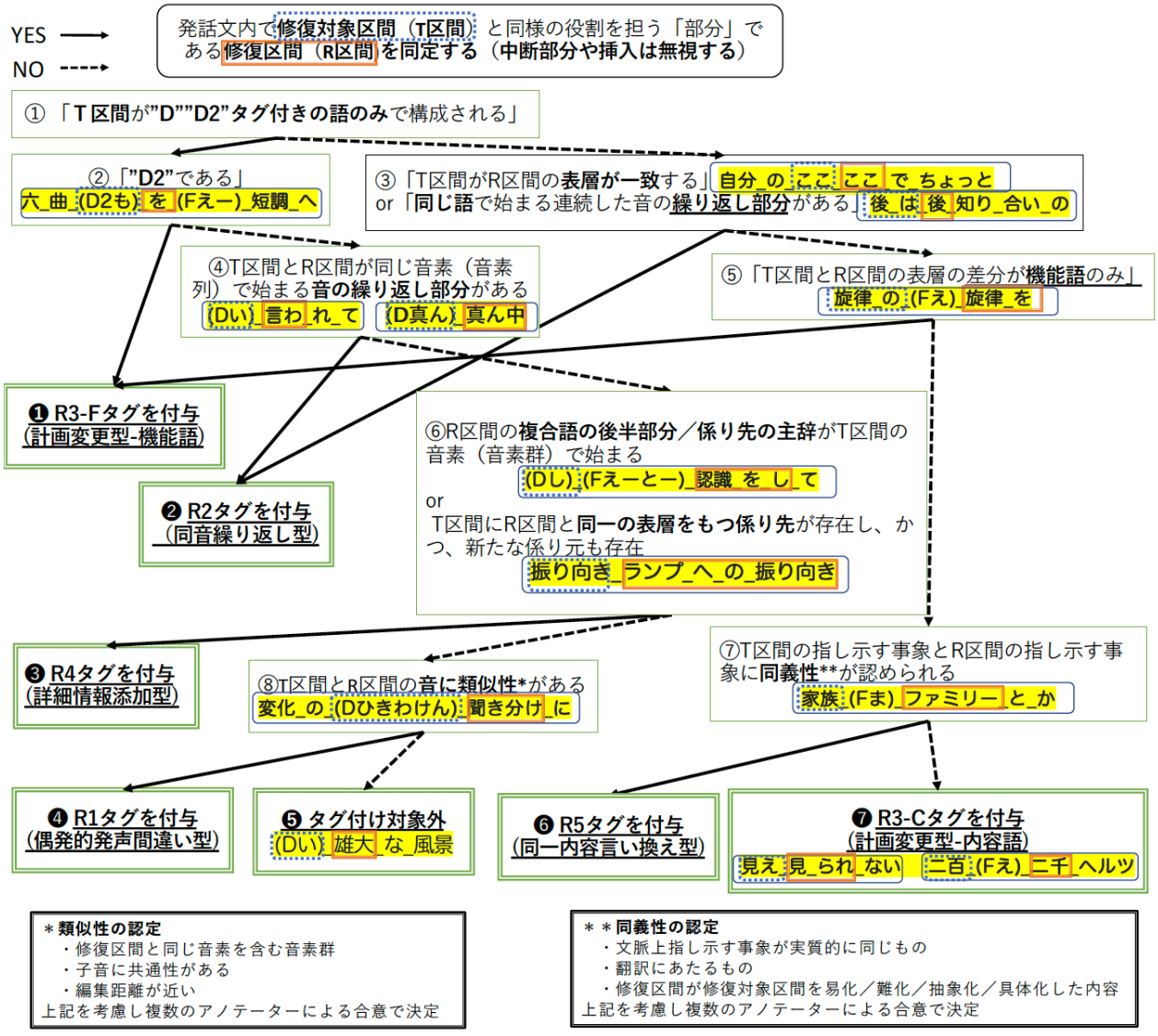


**図 1**　言い直しのタグを付与するための決定フローチャート