『日本語日常会話コーパス』に対する自然会話特有の現象を区別するための係り受け関係ラベルの付与

吉田 奈央 †　　宮尾 祐介 †
† 東京大学大学院情報理工学系研究科

1 はじめに

本研究は、自然な音声会話の構文解析の研究のため、『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation; CEJC) [1] に対して係り受け関係アノテーションデータを構築することを目的とする。特に、自然会話ではフィラー、言い直し、挿入、省略といった現象によって構文構造が崩れることに着目し、これらの現象を正確に表示するための係り受け関係ラベルを設計した。日本語の係り受けアノテーションデータはこれまで多数開発されており、ガイドラインの整備が行われている [2, 3]。しかし、書き言葉や独話を主として整備されてきたガイドラインは、自然会話のアノテーションにそのまま適用することは自明でない。

『日本語日常会話コーパス』は、国立国語研究所が開発を進めている大規模コーパスであり、さまざまな場面における自然な日常会話の音声・映像を収録したものである。本研究では、本コーパスのコアデータに対し、書き起こしテキスト、短単位レベルの形態論情報およびラベルなし係り受け関係がアノテーションされたデータを用いる。まず、本データの形態論情報・係り受けアノテーションを分析し、自然会話特有の現象のために通常の係り受け関係が付与できないケースを収集する。そして、これらの現象を分類し、各現象を区別するための係り受け関係ラベルを定義する。本稿では、10 会話 5480 発話文について分析・アノテーション作業を行った。以下では、アノテーションの作業方法と、開発したデータの分析結果を報告する。

2 先行研究

本節では、本論文でガイドラインを参照している日本語話し言葉コーパス (Corpus of Spontaneous Japanese; CSJ) および現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Written Japanese; BCCWJ) について詳述する。また、自然会話特有の言語現象に関する既存研究について説明する。

2.1 CSJ

日本語話し言葉コーパス（CSJ）[1)] [4] は、講演などの音声発話データを収集した大規模コーパスである。本データの一部に対し、文節間係り受け関係が付与されている [2]。CSJ は自由対話のデータも含むが、係り受け関係の付与対象となっている発話データはコアデータに含まれる独話とテストセット[2)]のみである。

CSJ では、音声発話特有の現象について「自発的な話し言葉特有の現象として、<中略>、言い差し（言いやめ）、言い直し、言い換え、挿入構造、倒置、ねじれなどの非流暢現象があげられる」と説明している [2]。フィラー、接続詞、感動詞、非言語音については、係り受け関係を付与せず、当該文節に対してそれぞれ個別のラベルを付与して区別している。言い直し・言い換えは他の係り受け関係と区別するラベルが付与されているが、言い直し・言い換えの種類については分類されていない。発話が途中で中断されるなど係り先がない場合は係り受け関係を付与せず、個別のラベルを付与している。

また、複数人の会話を対象としておらず、かつフォーマルな場での発話であるため、本研究が対象とする複数人の日常会話とは同じ音声発話であっても見られる現象に差があると考えられる。例えば、川田 [5] の CSJ のフィラー分析で説明されるように独話では対話よりフィラーが多く見られるという指摘がある。さらに、複数の参与者からなる会話では、話者交代によって生じる独話にはないインタラクションがある。例えば、ターン取りのために挟むフィラーや接続表現、他の参与者に邪魔されたため文末を持たないまま終了した発話文、発話文中に現

1） https://pj.ninjal.ac.jp/corpus_center/csj/
2） https://pj.ninjal.ac.jp/corpus_center/csj/manu-f/overview.pdf

れる他の参与者に対する呼びかけや短い応答などが頻出する。これらはそれぞれ区別して分析をされるべき現象だが、既存の係り受け関係アノテーションでは明確に弁別できない。

2.2 BCCWJ

現代日本語書き言葉均衡コーパス (BCCWJ)[3] [6] は、出版物及びインターネット上の書き言葉を収集した大規模コーパスである。コアデータに対して文節間係り受け関係が付与されており、アノテーションガイドラインが整備されている [3]。CSJ のアノテーションガイドラインをベースとしているが、CSJ で区別されているフィラーや接続詞等は区別しない。フィラーや文末が省略されている場合など、なんらかの理由で係り先がないあるいは認定しない文節は特別なノード ROOT を係り先として付与し、係り受け関係ラベル F を付与して通常の係り受け関係とは区別している。本研究では、本ガイドラインに基づいて付与された係り受け関係データを用いて分析を行う。

2.3 自然会話特有の現象

2.3.1 フィラー

川田 [5] によると、フィラーとは発話文に差し込まれる形で現れる、場繋ぎに使用される語で、概念的意味を持ち合わせない間投詞であるとされる。自然発話に現れることが古くから知られており、「談話における話し手の心的操作を表示するもの」であり「円滑な談話進行を可能にするもの」[7] として捉えられる。CSJ の転記基準においては、フィラーを「あー」「えー」「うーん」などと表記される類型パターンをもとに分類しフィラー表現をまとめており [8]、CEJC でもこれを採用している。よって本研究の対象データもこの基準によって認定されたフィラーを採用している。

2.3.2 挿入

挿入とは、ある文の途中に別の文もしくは別の構造を持つ要素が挟まる現象を指す。日本語の発話における挿入表現については、丸山 [9] が CSJ のフォーマルな場での独話を対象として挿入の談話内での機能に注目した分析を行っているが、係り受け関係は区別されていない。また、ここで判断基準とされている「挿入の前後に係り受け関係のある要素（短単位）が存在すること」「文末表現が接続助詞で終わるものであること」については、日常会話においては必ずしも成り立たない。

2.3.3 言い直しに類するもの

言い直しは、修復や不足の情報を補う目的で同義の語が繰り返されることである。言い換えは言い直しと同様の目的で、ほぼ同義であるが表層の違う語が当該部分の後ろに出現する現象である。これらは、同じ役割の語が一発話文に複数存在することになり文法構造を複雑にする一つの要因である。

丸山 [10] は言い直しと言い換えを言い直し表現として扱い、それらの機能的分析を行う中で構造を元に 5 種類のパターンに分類している。2.3.2 節と同様に CSJ の独話を対象として使用しており対話における言い直し表現は今後の課題としている。

3 データ概要

本研究では、形態論情報および文節間係り受け関係が付与された日本語日常会話コーパス (CEJC) コアデータ 52 会話のうち、200 発話文以上ある複数話者による日常会話 10 会話 5480 発話文を使用した。本データは現代日本語書き言葉均衡コーパス (BCCWJ) で用いられたアノテーションガイドライン [3] にもとづいており、通常の係り受けを表すラベル D と、フィラーなど係り受け関係を付与しないことを示すラベル F が付与されている。

本データに対し、本研究では以下の手順で分析およびアノテーションの修正・追加を行った。

1. 各会話データについて、各参与者の発話分をそれぞれ別ファイルに出力する（1 会話に 3 人の参与者がいれば、各々の発話のみが収められたファイルが 3 ファイルできる）。
2. 形態論情報を参照し、表 1 に示すラベル C, F, D, Int, E を文節に付与する。
3. ラベル C が付与された文節について、1 発話文中における係り受け関係を修正・追加する。

上記 3 では、アノテーション作業及び係り受け関係を可視化するツールとして brat[4] を使用した。

3） https://pj.ninjal.ac.jp/corpus_center/bccwj/

4） https://brat.nlplab.org/

表 1　本研究で用いるラベルセット

文節に付与されるラベル		付与数
C	一文節を示す。	12267
F	フィラーとして扱う一文節を示す。係り受け関係付与対象外。	301
D	短単位に満たない語断片を示す。係り受け関係付与対象外。	247
Int	フィラー以外の感動詞を示す。係り受け関係付与対象外。	28
E	笑い声など語にならない音を示す。係り受け関係付与対象外。	167
係り受け関係に付与されるラベル		**付与数**
D	通常の係り受け関係を示す。	5679
N	発話文外に係り先がある場合や、述語が省略されている場合など、当該発話文内に係り先がない場合を示す。	468
U	フィラー相当の語句など、文の命題に寄与しないため係り先がはっきりしない場合を示す。	113
J	文頭の接続詞で文と文をつなぐ役割をするものを示す。	269
I	挿入によって挿入前の要素の係り先が失われている場合を示す。	13
RS-S	言い直しにおいて言い直し元と言い直し先の表層が同じ場合を示す。[10] の R2 と R4 に相当。	113
RS-D	言い直しにおいて言い直し元と言い直し先の表層が異なる（別表現）場合を示す。[10] の R3 と R5 に相当。	37

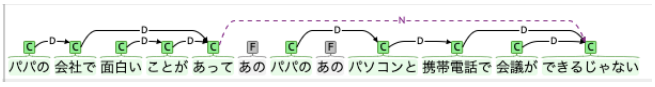

図 1　発話文中に係り先がない例（ラベル N）

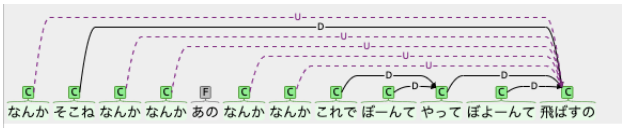

図 2　フィラー化した「なんか」（ラベル U）

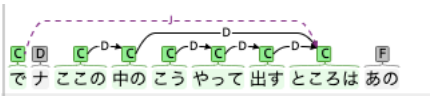

図 3　文頭の接続表現（ラベル J）

4　係り受け関係ラベルの付与

本節では、BCCWJ のガイドライン [3] に基づく係り受け関係アノテーションでは区別できない自然会話特有の現象について分析を行い、係り受け関係ラベルを付与することでこれらを正確に表示する方法を述べる。本研究で用いる係り受け関係ラベルセットと、当該データにおける付与数を表 1 に示す。

4.1　発話文中に係り先がない事例

自然会話では、当該の発話文外の要素に係り先が存在する、発話を途中で止めたことで係り先となる部分が消失する、といった理由によって当該発話文内に係り先がない事例がある（図 1）。BCCWJ の基準では、係り先がない文節は特別なノード ROOT を係り先とし、ラベル F を付与することで通常の係り受け関係と区別する。本研究では、当該発話文内に係り先がない場合は、便宜的に最右要素を係り先とし、ラベル N を付与する。これにより、通常の係り受け関係や後述する他の現象と区別されるため、構文構造や意味内容の分析を行う場合や、将来的に他の発話文内の係り先を認識する場合などにおいて要対応箇所として認定可能となる。

4.2　係り先が不明確な事例

副詞「なんか」「まあ」「ちょっと」などについては、CSJ のフィラー類型一覧には入っておらず、CEJC の形態論情報ではフィラーと認定されていない。しかし、これらが実質的にフィラーとして使用されている事例が見られる（図 2）。これは、「不確定な要素を述べる先行並行用法のようだが呼応する要素がはっきりしない例」、「一発話文に内に多発しフィラー化している『なんか』」[11, 12] に相当すると考えられる。したがって、これらの表現が発話文の命題に実質的に寄与していないと判断される場合は、最右要素を係り先とし、他の係り受け関係と区別するためにラベル U を付与する。

また、発話文の先頭におかれ、前後の発話文をつなぐ接続詞や接続助詞が頻出する（図 3）。これらは文と文とを自然に接続する役割を果たしていると考えられるため、その係り先は発話文の最右要素とする。ただし、自然会話においては文末が省略されるなど文が完成しない場合も多く、この場合当該接続詞・接続助詞は文法的には最右要素に係るわけではない。したがって、通常の係り受け関係と区別するため、この係り受け関係についてはラベル J を付与する。なお、CSJ では接続詞は係り受け付与対象外となっているが、本研究ではその他の接続詞については通常の係り受け関係を付与する。

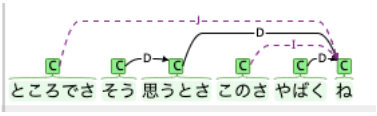

図 4　挿入により係り先が消失する例（ラベル I）

表 2　丸山による言い直し 5 分類と本研究のラベル

R1	発音誤り（語断片/発音ミス起因)	対象外
R2	単純な繰り返し (同一表層)	RS-S
R3	語の選択誤り（別表層の語句/助詞誤りを含む)	RS-D
R4	情報不足（新規の語が添加された同一表層)	RS-S
R5	別表現へ言い換え（同義の別表層の語句/カタカナ語の翻訳)	RS-D

固有名や一般名詞を使用した呼びかけ（「お母さん」など）については係り先のない語として 4.1 節と同様にラベル N を付与する。

4.3　挿入による係り先の消失

対象データにおいては、図 4 のように挿入部分で発話文が終わり、挿入文前に後続する発話が消失した例が見られる。これは挿入部の前と挿入部後に係り関係があるという丸山 [9] の定義には当てはまらないが、文法的な整合性を無視した発話部分が差し込まれており、挿入構造と判断するのが妥当と思われる現象が生じている。本研究では、その当該部分を除いても発話文全体の話脈に影響がない場合に当該部分を挿入部と同定し、後半部分がなくとも挿入構造と同等のものとして扱う。

また、この場合挿入部前の文節の係り先が失われているため、その係り先は最右要素とし、挿入によって係り先が消失したことを示すためにラベル I を付与する。挿入された部分は、挿入部内での係り受け関係を付与するが、係り先がない場合は最右要素を係り先とし、関係ラベルは N とする。

4.4　言い直し表現

係り受け関係を同定する際の言い直し表現の問題は、図 5,　図 6 でみられるように、発話文の命題において同じ役割をもつ要素が複数存在することである。対象データに置いても言い直し表現が多く見られる。これらの現象は、直前の発話が文法構造的に不完全な形で中断されること、中断後に同じ表層あるいは同義の語が発話されることから認定できる。

本研究では、丸山 [10] による言い直しの分類にもとづき、言い直された語の表層が同一かどうかという基準で 3 パターンに分類して係り受け関係ラベルを付与した（表 2)。R1 は、係り受け付与対象が語断片となるため、3 節に基づき対象外とする。被言い直し部と言い直し部が同一の表層を含む R2, R4 の場合、被言い直し部と言い直し部の間にラベル RS-S を付与する。R3, R5 は、被言い直し部と言い直し部の間にラベル RS-D を付与する。いずれの場合も、被言い直し部・言い直し部とその外との係り受け関係は、言い直し部に対して付与する。

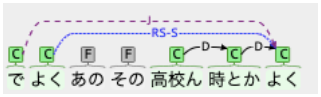

図 5　同じ表層の言い直し（ラベル RS-S）

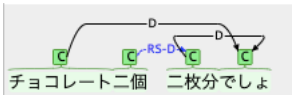

図 6　異なる表層の言い直し（ラベル RS-D）

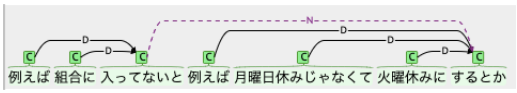

図 7　言い差し・文末要素の省略の例（ラベル N）

4.5　言い差し・文末要素の省略

言い差しおよび文末要素の省略における問題点は、文法的に不完全な中断により、係り先となる要素が消失していることである。これは 4.1 節と同様の現象であるとし、ラベル N を付与し、係り先は最右要素とした。

5　まとめ

本研究では、日本語日常会話コーパスの発話文に対し、係り受け関係の付与を行った。特に、フィラーや言い直しなど、通常の係り受け関係とは異なるものや、係り先が存在しない場合など、自然会話特有の現象について分析を行い、これらを区別できるように係り受け関係ラベルを設計・付与した。現時点では、問題となる現象の同定とそれを区別する手法を策定した段階である。今後は、本手法に基づき効率的かつ高品質なアノテーションを行うためにアノテーションガイドラインを整備し、データを拡張する予定である。

謝辞　本研究は国立国語研究所機関拠点型共同研究プロジェクト大規模日常会話コーパスに基づく話し言葉の多角的研究およびコーパス開発センター共同研究プロジェクトによるものです。

参考文献

[1] 小磯花絵, 天谷晴香, 石本祐一, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 伝康晴, 西川賢哉. 『日本語日常会話コーパス』モニター公開版コーパスの設計と特徴. 国立国語研究所「日常会話コーパス」プロジェクト報告書 3, 3 2019.

[2] 内元清貴, 丸山岳彦, 高梨克也, 井佐原均. 『日本語話し言葉コーパス』における係り受け構造付与.

[3] 浅原正幸, 松本裕治. 『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション. 自然言語処理, Vol. 25, No. 4, pp. 331–356, 2018.

[4] 国立国語研究所. 『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』. 2006.

[5] 川田拓也. 日本語フィラーの音声形式とその特徴について : 聞き手とのインタラクションの程度を指標として. PhD thesis, 京都大学, 2010.

[6] 国立国語研究所コーパス開発センター. 現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版.

[7] 田窪行則, 定延利之. 談話における心的操作モニター機構. 言語研究, Vol. 1995, No. 108, pp. 74–93, 1995.

[8] 小磯花絵, 西川賢哉, 間淵洋子. 報告書『日本語話し言葉コーパスの構築法』, 第二章　転記テキスト.

[9] 丸山岳彦. 『日本語話し言葉コーパス』に基づく挿入構造の機能的分析. 日本語文法, Vol. 14, No. 1, pp. 88–104, mar 2014.

[10] 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, sep 2008.

[11] 鈴木佳奈. 会話における「なんか」の機能に関する一考察. 大阪大学言語文化学, No. 9, pp. 63–78, 2000.

[12] 宮永愛子, 大浜るい子. 会話における「なんか」の働き–大学生による自由会話データを中心に. 表現研究, No. 91, pp. 30–40, mar 2010.