『日本語日常会話コーパス』に対する係り受け情報アノテーション

浅原正幸
国立国語研究所
masayu-a@ninjal.ac.jp

若狭絢
国立国語研究所

概要

音声言語に対する日本語係り受けツリーバンクは、独話を主とする『日本語話し言葉コーパス』に基づくものであった。対話を主とする『日本語日常会話コーパス』に対して係り受け情報を付与作業を進めており、2022 年春に公開予定である。本発表では国語研長単位形態論情報・文節境界・文節係り受けアノテーション作業の概要を示すとともに、2022 年 1 月時点での基礎統計を示す。

1 はじめに

国立国語研究所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」では、『日本語日常会話コーパス』(CEJC) の構築を進めており、50 時間分の会話データを 2018 年 12 月にモニター公開し [1]、2022 年春には 200 時間分の会話データを公開予定である。CEJC は設計 [2, 3] に基づき、収録 [4, 5]・転記 [6]・発話単位付与 [7, 8]・国語研短単位形態論情報付与・国語研長単位形態論情報付与・文節境界付与・文節係り受けアノテーションの順に作業を実施した。本発表では CEJC に対する長単位形態論情報付与・文節境界付与・文節係り受けアノテーションについて示す。

書き言葉を対象とした文節係り受けアノテーションとして、1995 年の毎日新聞記事を対象とした京都大学テキストコーパス [9]、京都大学ウェブリード文コーパス [10]、『現代日本語書き言葉均衡コーパス』(BCCWJ) を対象とした BCCWJ-DepPara [11] がある。話し言葉を対象とした文節係り受けアノテーションは『日本語話し言葉コーパス』[12] に含まれているが、主として独話を対象としており、対話を対象とした大規模な係り受けアノテーションは存在しない。

2 節では、アノテーション作業の概要として、国語研長単位形態論情報・文節境界アノテーション作業と文節係り受けアノテーションについて示す。3 節では、2022 年 1 月現在のデータの基礎統計を示す。

2 アノテーション作業

2.1 国語研長単位形態論情報・文節境界アノテーション

以下では国語研長単位形態論情報・文節境界アノテーションについて述べる。上流工程として収録・転記・短単位形態論情報が完了しているデータに対して、長単位解析器 Comainu [13] により解析したデータを対象に進める。短単位形態論情報付与 [14] では、転記に基づく形態に対する形態素の認定を行う。この時点では「可能性に基づく品詞体系」が付与される。長単位形態論情報付与 [15] では、文節の認定を行った上で、各文節の内部を規則に従って自立語部分と付属語部分に分割していくという手順で単位認定を行う。複合的な機能表現を含めて、複合語を構成要素に分割することなく全体で一つとして扱う。長単位の品詞認定においては、「用法に基づく品詞体系」として、複合化した結果の単位が文脈内で統語的にどのように振る舞うかに基づき、品詞を認定する。作業は 2019 年 2 月から 2020 年 8 月にかけて、作業者 2 名[1)] により、2 次チェックまで行った。その後は、上流工程の修正が行われた際に、もしくは、係り受け情報アノテーションで不都合な点が見られた際に、追随して長単位形態論情報に対する修正を繰り返し行った。作業には国語研内のコーパス管理システム『大納言』[16] を用いた。

単位認定基準は基本的に BCCWJ で整備された規程集 [15] に基づく。しかしながら、話し言葉を付与

1） 短単位形態論情報修正を作業した 3 名が、一時的に作業を補助したこともあった。

するうえで規程に定義されていない表現が出現した。そのような場合、**より自然な係り受け木を表現しやすい単位**を認定することを心がけた。

以下具体的な例を示す。

例えば、言いよどみが長単位の中に包摂したほうがわかりやすい場合には包摂するようにした。

(1) || 店員 | シ | さん || が ||
【T001_009-IC03】

短単位境界を |、長単位境界を || で表現する。(1) では、言いよどみ「シ」を含む形で、長単位「店員シさん」を認定する。

他の例として省略による長単位解析誤りがある。書き言葉においては助詞が手がかりとなり文節境界が認定しやすい。

(2) || 余裕 || なかっ | た || 気 | が || する ||
【T001_009-IC01】

(2) では、「余裕」のあとの助詞が省略されているために長単位解析器が誤って連結していたのを修正した。

(3) || 結構 || し | てる || 人 || いる | よ ||
【T006_008a-IC04】

(3) では、「結構する」という単位が誤解析されていたが、分割した。

長単位作業時に短単位作業の誤りが多く見つかった例として、フィラーと代名詞・連体詞の曖昧性がある。

(4) || あれ || もう || 声変わり ||
【T011_007-IC06】

(4) では、「あれ」が当初代名詞とされていたが、音声を聞きながら係り受け・共参照を含めて検討すると指示するものもない表現であったため感動詞とした。

(5) || 酵素 | うん | たら | かん | たら || って ||
【T002_019-IC01】

(5) では、短単位の規程上「云（名詞）」「たら（副助詞）」「かん（代名詞）」「たら（副助詞）」と分割されているが、係り受けを付与するにあたって「酵素うんたらかんたら」の 1 普通名詞に結合した。

2.2 文節係り受けアノテーション

係り受けアノテーションは BCCWJ-DepPara[11] の基準に準じる。BCCWJ-DepPara では、通常の係り受け相当の "D"、文境界相当の"Z"、係り受けを付与するうえで後続文節と連結する"B"、係り先が決められない "F" の 4 つのラベルを認定している[2)]。この 4 種のラベルを本研究にも適用するが、**係り受けが決められない文節にはラベル"F"を付与し、無理に係り受けを決めない**という基本方針を決め、作業を進めた。BCCWJ-DepPara では並列構造の範囲についても付与したが、CEJC に対しては現在のところ並列構造の範囲の付与を行っていない。作業にはコーパス管理システム『ChaKi.NET』[17] を用いた。作業は 2020 年 4 月から開始し、2021 年 3 月に 2 次チェックまで終了した。その後は、上流工程の修正が行われた際に、追随して係り受け情報に対する修正を繰り返し行った。

アノテーションする単位は、発話単位 [7] に基づく。このため、書き言葉の文とは異なり、返答など 1 文節 1 発話単位のものが多い一方、明確に文末が定義できない長い発話単位もあり、単位末が終助詞・接続助詞で終わることも多い。また、単位を 1 発話単位としてするため、話者間の係り受けは付与しない。

図 1 に CEJC に対する係り受けアノテーション例を示す。フィラーや言いよどみが 1 文節をなす場合（図中 1 例目と 2 例目）には、それぞれラベル"F"を付与したうえで、最後の要素に係るようにした。また、発話末は必ずしも述語であるわけではない。図中 3 例目のように、発話末が副詞のような述語に相当しない場合もある。その場合、述語に係るべき要素は最終要素に係るようにした。

3 基礎統計

2022 年 1 月現在、上流工程の転記・短単位の修正がまだ続いており、最終の基礎統計は異なる可能性がある。以下では、暫定版として 2022 年 1 月 10 日時点の情報を示す。

文節数	134,027 文節
発話単位数	58,175 発話
平均発話長	2.30 文節
係り受け関係数	75,852
発話末に係る文節数	46,173 文節

表 1 基礎統計

2） CSJ[12] では、通常の係り受けをラベルなしとし、並列"P"・部分並列"I"・狭義の同格"A"・広義の同格（総称など）"A2"・倒置"R"・係り受けの交差"X"・接続詞"C"・感動詞"E"・言い直し"D"・フィラー"F"・係り先のない文節"N"・呼びかけ"Y"などのラベルが定義されている。

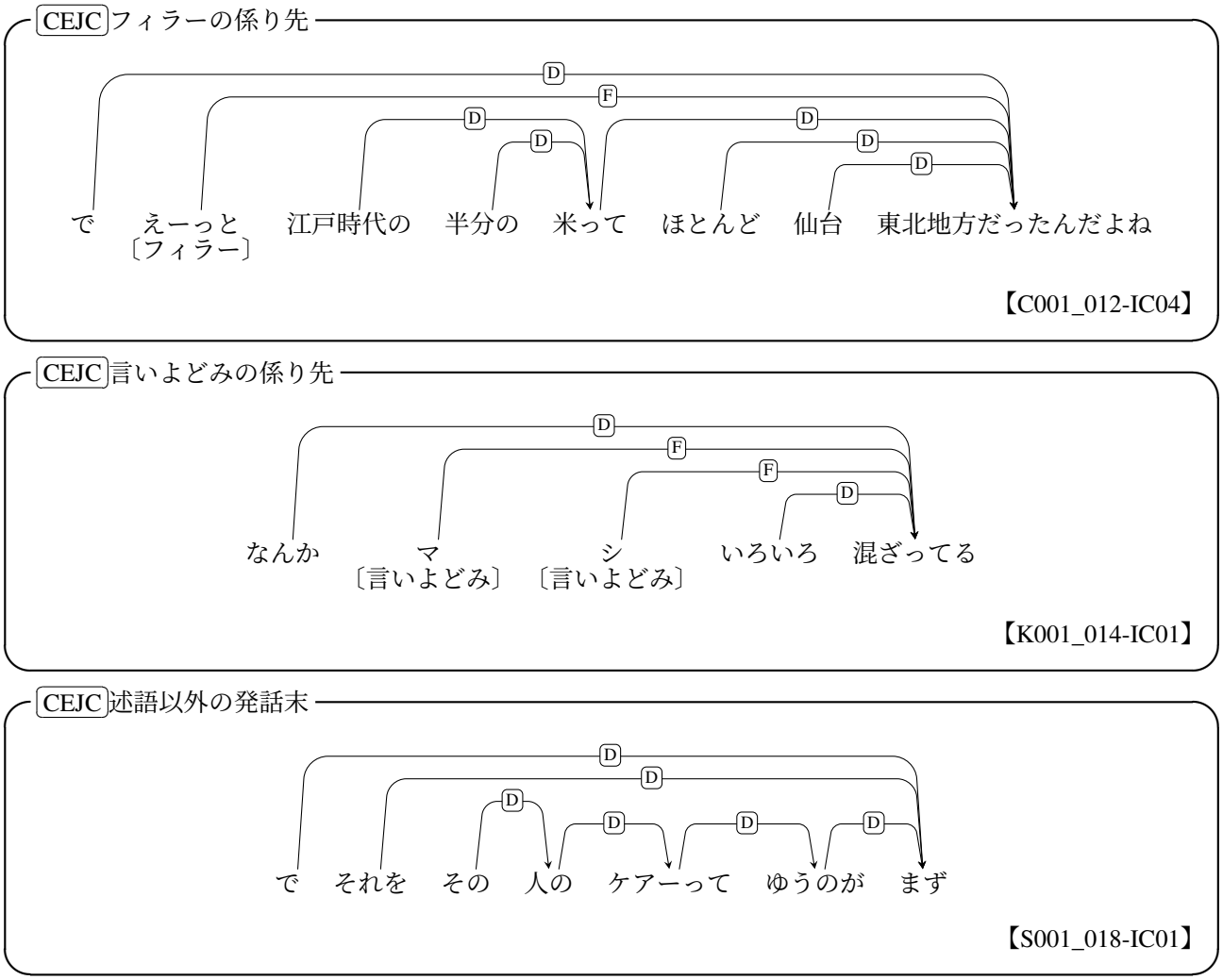

図 1　CEJC に対する係り受けアノテーション例

表 1 に基礎統計を示す。平均発話長を 2.30 文節と書き言葉と比べて極端に短い。発話末に係る文節数が発話単位数より短いのは、後に示すとおり 1 文節 1 発話単位のものが多いためである。また全係り受け数 75,852 のうち 46,173 (60.87%) が発話末に係る（右隣接要素に係るものも含む）。

ラベル	頻度	割合
D	126,162 文節	94.13%
F	7,561 文節	5.65%
Z	304 文節	0.22%

表 2　係り受けラベルの分布

表 2 に係り受けラベルの分布を示す。94.13%の文節が通常の係り受けラベルで、5.63% の文節がフィラー・言いよどみなどの係り受けが決められないものであった。BCCWJ-DepPara においては、長単位作業者（国語研内）と係り受け作業者（奈良先端大内）とで異なる作業者が従事していたために、その齟齬を吸収するために "B"ラベルを利用したが、今回の作業者は長単位と係り受けの両方の作業に従事しているために、このラベルは不要であった。また、発話単位に基づくために、「文の入れ子」を表現するための "Z"も 0.22%と少なかった。

表 3 に発話長（１発話内の文節数）の分布を示す。58,174 発話中 34,340 発話が 1 文節の発話、8,779 発話が 2 文節の発話（合わせて 74.12%）であった。2 文節以下の場合、特殊な場合（係り先がない・2 文に分かれる・倒置など）を除いて、係り受け木の構造は一意に決まるため、実質的なアノテーション作業は発生しない。係り受けアノテーション作業は残りの 3 文節以上の発話が対象となるが、2 文節以内のものについても長単位の再定義・フィラー・言いよどみの認定などに時間を要し、場合によって、短単位形態論情報の修正を行った。

発話長	頻度
1	34,340
2	8,779
3	5,184
4	3,076
5	1,960
6	1,315
7	893
8	654
9	450
10	352
11	230
12	207
13	152
14	114
15	95
16	77
17	59
18	49
19	32
20	31
21	16
22	25
23	15
24	13
25	7
26	12
27	6
28	5
29	4
30	4
31	3
32	5
33	3
34	2
35	3
36	1
45	1

表 3　発話長の分布

係り受け長	頻度
-4	1
-3	2
-2	2
-1	8
1	44191
2	12391
3	6775
4	4087
5	2551
6	1605
7	1129
8	776
9	543
10	396
11	316
12	219
13	182
14	157
15	117
16	80
17	69
18	49
19	39
20	22
21	26
22	24
23	15
24	16
25	15
26	6
27	6
28	6
29	3
30	8
31	8
32	6
33	2
34	2
35	1
43	1

表 4　係り受け長の分布

表 4 に係り受け長の分布を示す。全部で 75,852 の係り受け関係のうち 44,491 (58.65%) のものが右隣りのものにかかる（係り受け長 1）であった。係り受け長が負の値のものは倒置の係り受けであることを表す。倒置の係り受けが少ないのは、倒置が発生した箇所で発話単位が分割されて、別の発話単位に認定される傾向があるためである。

4　おわりに

本発表では、『日本語日常会話コーパス』(CEJC) に対する係り受け情報アノテーションについて解説した。2022 年春に有償版 CEJC 契約者向けに共有する予定である。

本研究は CEJC のコアを対象に、明確な係り受け関係を "D" ラベルで付与するとともに、不明確な係り受け関係を "F" ラベルで付与したに過ぎない。より精緻な話し言葉向けの係り受け関係ラベルが、吉田ら [18] により提案されている。本データに基づいて、話し言葉特有の不明確な係り受け関係のラベル付けが進むことを期待する。

また統語的な性質としては、省略が多く、発話単位に分割されているため、統語分析に向かないデータとなった。実質的な統語分析を行うためには、名詞句省略のみならず、述語省略を含めたゼロ照応を定義した述語項構造のアノテーションが必要であろう。その作業においては、述語項構造に基づく共参照情報を付与することが重要である。共参照情報についても、本データの場合「同一話者の異なる発話(単位)」「異なる話者の発話」「発話されないが動画に写り込んでいるモノ（コト)」「発話されず、動画にも写り込んでいないモノ（コト)」といったレベルがある。述語項構造のアノテーションに向けて、名詞句省略・述語省略を含めた共参照情報の整理を進めたい。

謝辞

実アノテーション作業に従事した山下華代氏に感謝の意を表します。

本文中にもお示しした通り、本研究実施時に東京大学の吉田奈央氏・宮尾祐介氏に様々な助言を受けました。その助言の一部は昨年の NLP2021 で発表されております [18]。ここに記して、感謝の意を表します。

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトの成果です。また、科研費 17H00917 の支援を受けました。

参考文献

[1] 小磯花絵, 天谷晴香, 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』モニター版の設計・評価・予備的分析. 国立国語研究所論集, No. 18, pp. 17–33, jan 2020.

[2] 小磯花絵, 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』の構築. 言語処理学会第 23 回年次大会発表論文集, pp. 775–778, 2017.

[3] Hanae Koiso, Yasuharu Den, Yuriko Iseki, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, and Yasuyuki Usuda. Construction of the Corpus of Everyday Japanese Conversation: An interim report. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. European Language Resources Association (ELRA), May 2018.

[4] 田中弥生, 柏野和佳子, 角田ゆかり, 伝康晴, 小磯花絵. 『日本語日常会話コーパス』の構築：会話収録法に着目して. 国立国語研究所論集, No. 14, pp. 275–292, Jan 2018.

[5] 田中弥生, 柏野和佳子, 角田ゆかり, 伝康晴, 小磯花絵. 『日本語日常会話コーパス』の構築–個人密着法に基づく会話の収録–. Technical report, 国立国語研究所プロジェクト報告書, March 2018.

[6] 臼田泰如, 川端良子, 西川賢也, 石本祐一, 小磯花絵. 『日本語日常会話コーパス』における転記の基準と作成手法. 国立国語研究所論集, No. 15, pp. 177–193, July 2018.

[7] Japanese Discourse Research Initiative. 発話単位ラベリングマニュアル version 2.1. Technical report, Japanese Discourse Research Initiative, Jan 2017.

[8] Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In **Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)**, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[9] Sadao Kurohashi and Makoto Nagao. Buinding a japanese parsed corpus while improving the parsing system. In **Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-98)**, pp. 719–724. European Language Resources Association (ELRA), 1998.

[10] 萩行正嗣, 河原大輔, 黒橋禎夫. 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析. 自然言語処理, Vol. 21, No. 2, pp. 213–248, 2014.

[11] 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.

[12] 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 (version 1.0). Technical report, 『日本語話し言葉コーパス』の解説文書, 2003.

[13] 小澤俊介, 内元清貴, 伝康晴. 長単位解析器の異なる品詞体系への適用. 自然言語処理, Vol. 21, No. 2, pp. 379–401, 2014.

[14] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（下）. Technical report, 国立国語研究所内部報告書, 2011.

[15] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版（上）. Technical report, 国立国語研究所内部報告書, 2011.

[16] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用. 自然言語処理, Vol. 21, No. 2, pp. 301–332, 2020.

[17] Masayuki Asahara, Yuji Matsumoto, and Toshio Morita. Demonstration of ChaKi.NET – beyond the corpus search system. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations**, pp. 49–53, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[18] 吉田奈央. 『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. 言語処理学会第 27 回年次大会発表論文集, pp. 1129–1133, 2021.