発話の流暢性を踏まえた機能表現の抽出と分析

土屋 智行 † 伝 康晴 ‡† 小磯 花絵 †
† 国立国語研究所 ‡ 千葉大学文学部
ttsuchiya@ninjal.ac.jp

1 はじめに

文を構成する要素が互いにどのような関係性を持っているのかを具体的に構造化するものが係り受けであるが，係り受けを構成する文節および文節同士の関係性は，必ずしも母語話者の直感や，意味的な関係性を直接反映するものではない．[1] では，隣接する文節をまたぎ，全体で機能的な意味をなす形態素の結合単位を「機能表現」とし，分析をおこなっている．[1] は，「2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現」を機能表現として取り上げ，形態素と係り受けという 2 つのレベルでの調整の必要性を述べている．

この機能表現は，文節をまたぐことに加え，文節と文節の間に入り込んだり，また複数の係り元を持つ文節を埋め込み節化することで，係り受け構造の複雑さや言語使用者の感覚との乖離を引き起こしうる．たとえば，図 1 のような係り受け関係は，機能表現を考慮していないために「保護者」と「参加する」という 2 つの文節の関係性を直接的に反映していない．また，図 2 のような係り受け関係の場合，命題の述語として複数の係り元を集約している文節「喜ぶんじゃないか」が,「と思う」に係っている．これによって，本来の発話内容の中心である命題内容や話者の推測が埋め込み節となり，情報の重要性が直接的に反映されない．

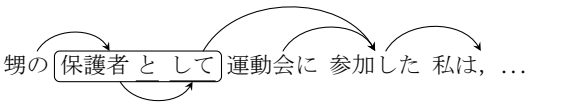

図 1: 係り受け関係を間接化する機能表現 (cf. [1])

機能表現の抽出や収集にあたって問題となるのは，文節というひとつの単位をまたいで存在するという特徴である．複数の構成要素の結合表現には，慣用句や連語などが挙げられるが，それらはコーパス上の生起状況によって一定の範囲まで抽出可能であるものの，構成要素の結合以外の基準の必要性も議論されている

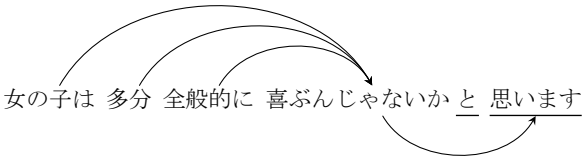

図 2: 複数の係り元を持つ文節を埋め込み節化する機能表現

[2]．文節に拠らない結合表現を機能表現として判断し，抽出する場合も同様の問題をはらんでいると言える．実際に，機能表現の用例集やデータベースの構築はなされている [3] ものの，機能表現やそれに近いカテゴリーの形態的な基準や範囲は，[4] においても明確に定められているものではない．

形態素の結合の度合いを可能な限り直接的に反映しながら機能表現を抽出するには，実際の話者の発話状況を観察する必要があるだろう．複数の構成要素の結合によって形成されるいわば「定型的」な言語表現によって，話者の流暢性が実現されるということは，理論的にも主張されている [5, 6]．また，言語使用者の感覚との乖離を埋めるためにも，話者の実際の発話状況から抽出する必要がある．

本論文では，文節に拠らない結合単位を抽出する指標として音声的な流暢性に注目し，コーパス全体での生起状況と音声的特徴の 2 側面から機能表現を分析する．まず話し言葉コーパスにおいて複雑な係り受けをなす発話から機能表現の候補を抽出した後，それらの機能表現として実現しうる単位を話し言葉コーパス上の頻度と発話内のポーズの特徴から分析していく．

2 方法

分析データは『日本語話し言葉コーパス』(第 3 刷) の RDB 版 [7] を使用した．まず事前調査として，図 3 のように，複数の係り元を持ち，かつ節末の係り先に単独で接続する文節を収集した．その結果，表 1 の

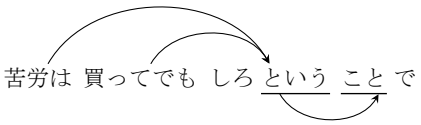

図 3: 収集の対象とした係り受け構造（下線は機能表現の箇所）

表 1: 複雑な係り受け関係にかかわる表現の分布

表現	# tokens
{ と/って }X	843
{ と/って } いう X	224
ように X	42
ような X	30
その他	919
全体	2058

ような結果となった．全体のうち，頻度が上位 4 位までのものは,「と X」「という X」「ように X」「ような X」であり，全体の半数を占めていた．その中でも,「{ と/って }X」の形式が最も多く，843 例であった．

次に，機能表現の候補となった上記 4 つの表現（以下 “KEY”）にたいして，今度は対象とする表現（{ と/って }, { と/って } いう，ように，ような）とその直後の語が異なる文節として係り受け関係にある表現全てを収集した．したがって，“KEY” は必ず直後の文節に係り，間に別の文節が挿入されるものは分析の対象外とした．さらに，分析に用いる変数として (1)〜(3) の情報を収集した．係り受け関係およびそれぞれの変数同士の関係は図 4 のように示される．なお,「と思う」「とする」「ようにする」などの機能表現を取り扱うため,「と X」「ように X」の X にあたる単語の品詞は，動詞に限定した．

(1) “KEY” の直前の長単位の語彙素（LeftLemma）

(2) “KEY” の直後の長単位の語彙素（RightLemma）

(3) (1) および (2) と “KEY” の間の（0.1 秒以上の）ポーズの有無（PauseL, PauseR）

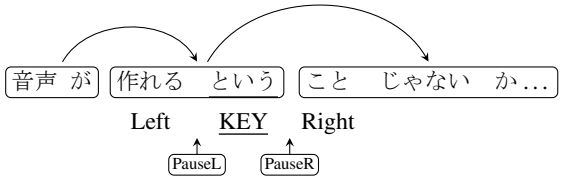

図 4: 機能表現の候補および収集する前後の長単位

3 結果

3.1 表現ごとの事例数と係り先

まず，それぞれの表現ごとに収集した事例数および上位 5，頻度 5 以上の直後長単位語彙素を表 2 に示す．

表 2: 表現ごとの事例数と上位の直後長単位語彙素

KEY	# tokens	直後長単位語彙素（上位 5, 頻度 5 以上）
{ と/って } いう	2844	事，風，感じ，形，話
{ と/って }	1525	言う，為る，思う，成る，考える
ような	813	事，形，感じ，気，結果
ように	168	成る，為る，読める
total	5350	

それぞれの表現の順位は，最初の予備調査と異なり,「{ と/って } いう X」が最も多く，それに続いて「{ と/って }X」「ような X」「ように X」という順番となった．またこれらの表現が係っている直後長単位の語彙素の上位は，手前の文節と係り受け関係を持つことで，全体として機能的な意味を示す表現，すなわち機能表現となる表現が上位を占めた．

3.2 直後長単位の語彙素とポーズ

次に，それぞれの機能表現の音声的特徴について見ていく．候補となった表現ごとの直後長単位の語彙素と，その間に（0.1 秒以上の）ポーズがある率を，表 3，表 4，表 5，表 6 に示した．

全体を見ると,「{ と/って }」「{ と/って } いう」「ように」では直後長単位との間にポーズをおかない傾向がみられた．特に「{ と/って }」「{ と/って } いう」は，直後長単位との間にポーズを置かずに発話している傾向が強かった．これに対して,「ような」は，全体では 15%がポーズを置いていた．

「{ と/って }」が接続する上位の語彙素のうち,「と言う」「と為る」「と成る」は特にポーズを置かない傾向が見られた．それにたいして,「と思う」「と考える」はポーズを比較的多く置く傾向があり，他の表現との差異が見られた．

「{ と/って } いう」では,「という事」「という風」「という形」がポーズを置く割合が低く,「という話」が最もポーズを置く割合が高かった.「という感じ」は,「という事」「という風」「という形」と同じようにポーズを置く率が低い傾向にあったが，表現全体における割

表 3: 「{ と/って }」の直後長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
言う	426	12	2.82%
為る	295	7	2.37%
思う	252	18	7.14%
成る	109	2	1.83%
考える	49	5	10.20%
その他	395	59	14.94%
全体	1526	103	6.75%

表 4: 「{ と/って } いう」の直後長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
事	1303	34	2.61%
風	404	5	1.24%
感じ	80	6	7.50%
形	44	2	4.55%
話	42	7	16.67%
その他	971	121	12.45%
全体	2844	175	6.15%

表 5: 「ような」の直後長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
事	115	10	8.70%
形	52	4	7.69%
感じ	46	13	28.26%
気	42	4	9.52%
結果	18	3	16.67%
その他	540	93	17.19%
全体	813	127	15.60%

表 6: 「ように」の直後長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
成る	50	3	6.00%
為る	46	4	8.70%
読める	9	0	0.00%
その他	64	7	10.94%
全体	169	14	8.28%

合よりも高く，平均してポーズが置かれている傾向にあると判断される．

「ような」では,「ような形」が最もポーズを置く割合が低かった．それに続いて「ような事」「ような気」もポーズを置かない傾向にあった．上位の表現の中でも,「ような感じ」がとりわけポーズを置く傾向にあり，全体の 3 割近くがポーズを置いていた．

「ように」は，頻度 5 以上で出現するものが「成る」「為る」「読める」のみだったため，抽出の対象となる言語表現は「ように成る」「ように為る」「ように読める」の 3 つだけであった.「ように成る」「ように為る」の中でポーズを置くものは観察できたが,「ように読める」でポーズを置く例は見られなかった．

3.3 直前長単位の語彙素とポーズ

次に，表現の直前に生起する長単位のポーズの置かれ方を表 7，表 8，表 9，表 10 に示す．直前の長単位におけるポーズの置かれ方は，直後の場合とは異なり,「ような」「ように」のほうがポーズが置かれる割合が低かった．表にもあるように，ポーズが置かれるものは，いずれの表現においても全体の 1 割程度であったが,「{ と/って } いう」では，1 割以上にポーズが置かれていた．

それぞれの表現の詳細をみていく.「{ と/って }」では，ポーズが置かれていない表現は意思や推量の意味を持つ「う」や断定の「だ」であった.「{ と/って } いう」では，終助詞「な」の後にポーズが置かれる割合が低かった.「ような」は，全体的に直前の長単位との間のポーズが少なく,「此のような」「というような」「何（ど）のような」「たような」で結合している傾向にある.「ように」も同様に直前の長単位との間のポーズが少なかった.「何（ど）のように」「のように」「此のように」「ないように」「同じように」という表現は全てポーズが置かれていなかった．

4 議論

1 節を踏まえ，3 節では文節をまたぎながらも高頻度で共起する流暢性の高い長単位の列を抽出した．その結果，発話中で機能的な意味を担うような言語表現がみられた．機能表現の抽出にあたって，話者が流暢に発話しているか否かという点を考慮することは妥当性が高いとここでは考える．具体的には,「{ と/って }」「{ と/って } いう」と隣接している長単位と結合している表現が，頻度および流暢性の観点からも機能表現として妥当なものが多いと考えられる.「{ と/って }」では「と言う」「と為る」「と成る」,「{ と/って } いう」では「という事」「という風」「という形」がそれにあたる．これらの表現は，機能的意味の側面でも，機能表現のカテゴリーに属するといえる．対して,「ような」における「ような形」,「ように」における「ように成る」「ように為る」は，他の「ような」「ように」の表

表 7: 「{ と/って }」の直前長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
か	90	10	11.11%
う	62	1	1.61%
な	57	4	7.02%
た	52	6	11.54%
だ	45	1	2.22%
その他	1220	119	9.75%
全体	1526	141	9.24%

表 8: 「{ と/って } いう」の直前長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
か	284	53	18.66%
た	185	35	18.92%
な	128	8	6.25%
ている	110	24	21.82%
ない	92	10	10.87%
その他	2046	323	15.79%
全体	2845	453	15.92%

表 9: 「ような」の直前長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
という	183	1	0.55%
此の	118	0	0.00%
た	71	1	1.41%
の	61	2	3.28%
何の	41	0	0.00%
その他	340	14	4.12%
全体	814	18	2.21%

表 10: 「ように」の直前長単位とポーズ率

語彙素	# tokens	# pauses	ポーズ率
何の	27	0	0.00%
の	19	0	0.00%
此の	19	0	0.00%
ない	9	0	0.00%
同じ	9	0	0.00%
その他	86	1	1.16%
全体	169	1	0.59%

現よりも機能表現に近いとも考えられるが，これについては今後検討が必要である．

対して，直後の長単位との間のポーズ率が比較的高かった表現，すなわち流暢性が相対的に低かった表現に「と思う」「という感じ」「という話」「ような気」「ような感じ」「ように為る」が挙げられるが,「思う」「感じ」「気」「話」という思考や伝達にかかわる語にそのような傾向が見られる点についても，今後詳細に分析する必要がある．

また，直前の長単位とのポーズ率を見ると,「ような」「ように」は，直後の長単位との結合が低かった代わりに直前の長単位との結合の度合いが高いことが確認できる．具体的には,「ような」の場合は「此のような」「というような」「何（ど）のような」「たような」,「ように」の場合は「何（ど）のように」「のように」「此のように」という単位でまとまって発話されている傾向が見られた．

機能表現は，係り受け関係を形作る文節の範囲をまたがって結合しているが，これは係り受け構造と機能表現のセグメンテーションのスコープが一致していない状態といえる [8]．発話上の音声的な流暢性を軸に据えて見直すことで，この係り受けと機能表現のミスマッチを一定の範囲まで収容することが可能となるだろう．

謝辞　本研究は国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴）による成果である。

参考文献

[1] 土屋雅稔, 注連隆夫, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一. 機能表現を考慮した日本語係り受け解析器学習のためのコーパス作成. 言語処理学会第 13 回年次大会論文集, pp. 510–513, 2007.

[2] Phoebe Ming Sum Lin. The phonology of formulaic sequences: A review. *Perspectives on formulaic language: Acquisition and communication. Continuum*, pp. 174–93, 2010.

[3] 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史. 日本語機能表現の自動検出と統計的係り受け解析への応用. 自然言語処理, Vol. 14, No. 5, pp. 167–197, 2007.

[4] 国立国語研究所. 現代語複合辞用例集. 2001.

[5] Charles J Fillmore. On fluency. *Individual differences in language ability and language behavior*, pp. 85–101, 1979.

[6] Andrew Pawley and Frances Hodgetts Syder. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and communication*, Vol. 191, p. 225, 1983.

[7] 小磯花絵, 伝康晴, 前川喜久雄. 『日本語話し言葉コーパス』RDB の構築. 第 1 回コーパス日本語学ワークショップ予稿集, pp. 355–364, 2012.

[8] 定延利之. ミスマッチを収容できる言語観を求めて. 音声文法研究会（編）, 文法と音声, pp. 167–196. くろしお出版, 1997.