韻律情報にもとづいた機能表現の抽出

土屋 智行（国立国語研究所言語資源研究系）†
伝 康晴（千葉大学文学部/国立国語研究所言語資源研究系）
小磯 花絵（国立国語研究所理論・構造研究系）

Extraction of Functional Expressions through Prosodic Information

Tomoyuki Tsuchiya (Dept. Corpus Studies, NINJAL)
Yasuharu Den (Faculty of Letters, Chiba University/Dept. Corpus Studies, NINJAL)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

要旨

機能表現には，複数の語がひとつのまとまりをなしているものや，文節の境界をまたぐものが多く存在する．したがって，語や文節という基準だけでは，機能表現を規定する形式的特徴を十分に抽出することができない．特に，文節の境界をまたぐ機能表現の場合，文の係り受け構造にも影響をあたえるため，結果的に言語使用者による文の解釈との乖離が生じてしまう．そこで，本発表では，言語使用者の解釈を反映する情報として母語話者が発話する際のポーズと韻律句という 2 つの韻律情報を用いて，文節をまたぎながらも，話者がひとつのまとまりとして発話している機能表現の抽出を試みる．具体的には，日本語話し言葉コーパス（CSJ）から隣接する 2 文節の係り受け関係やポーズの有無，韻律情報などの文節間の情報を抽出し，文節内における同様の情報との比較をとおして，文節間でひとつのまとまりをなしている表現を抽出し，その機能的な意味を考察する．

1. はじめに

文の構造は，係り受けによる文節の関係によって記述される．しかし，係り受けを構成する文節および文節同士の関係性は，自立語と付属語という形態論的な規定や，統語的な規定に基いているため，母語話者の直感や，意味的な関係性を直接反映しているものとはいえない．機能表現あるいは複合辞と呼ばれる言語表現は，複数の語が結合することでひとつの機能的な意味を有し，文の意味的な記述に重要な役割を持つものであるが，この機能表現と文節は，しばしばその境界に不一致がみられる．土屋ほか (2007) 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，機能表現を反映させた係り受け解析を試みている．その中で，土屋ほか (2007) は，機能表現の特定にあたって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．

上記のとおり，この機能表現の抽出や収集にあたって問題となるのは，複数の形態素が結合しているという特徴と，文節というひとつの単位に拠らないという特徴である．これまでの研

† ttsuchiya@ninjal.ac.jp

究 (国立国語研究所 2001, 土屋ほか 2007) では，一定の意味的な特徴に基づいて機能表現を収集しているが，形態的な基準が明確ではないために，特定の形態素の結合を機能表現と定めることが難しいことが指摘されている．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 (注連ほか 2007) 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，国立国語研究所 (2001) においても明確に定められているものではない．したがって，文節に拠らない結合表現を機能表現として判断し，抽出するためには，その形態的な特徴を分析していく必要がある．

では，どのような形態的特徴から 形態素の結合を分析していくべきだろうか．本論文では，実際の話者の発話状況を観察することで，形態素の結合の度合いを分析する必要があると考える．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 (Fillmore 1979, Pawley and Syder 1983)．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．

土屋ほか (2014) は，この流暢性に注目し，文節間のポーズ率に焦点を当てて，複雑な係り受け構造をもたらす機能表現の抽出を試みた．その結果，表 1 のような表現を機能表現の例として抽出した．

表 1: 土屋ほか (2014) が抽出した機能表現

係り元文節末語彙素	係り先文節先頭語彙素
{ と/って } いう	事，風，感じ，形，話
{ と/って }	言う，為る，思う，成る，考える
ような	事，形，感じ，気，結果
ように	成る，為る，読める

土屋ほか (2014) で挙げられた機能表現は，形式的な語彙（「事」「形」「為る」「成る」など）に加え，思考や伝達にかかわる語彙（「言う」「思う」「感じ」「話」など）が多用されている例が中心であった．しかし，発話における流暢性には，語間のポーズのみならず，韻律的なまとまりもかかわる．話者がひとつのまとまりとして発話している表現を抽出するには，韻律にも焦点を当てた分析が必要である．

本論文では，係り受け関係にある隣接した文節にたいし，文節という区切りを越えて結合している言語表現の抽出を試みる．その抽出の指標として，音声的な流暢性を取り上げる．ただし，その流暢性の基準には，具体的には，土屋ほか (2014) で使用されたポーズの出現の度合いに加え，韻律に基づいたまとまりとしてアクセント句を流暢性の指標として用い，分析していく．

2. 分析

2.1 方法

分析データは『日本語話し言葉コーパス（CSJ）』（第 3 刷）の RDB 版 (小磯ほか 2012) を使用した．まず，2 つの文節が係り受け関係にあり，かつ隣接している箇所を収集した．次に，収集した箇所から，文節間の流暢性を測る指標として，

- 文節境界およびその前後の長単位間のポーズ情報
- 文節中に出現するアクセント句境界情報

の 2 点を収集した．

文節境界およびその前後の長単位間のポーズ情報としては，係り元文節の末尾から 3 つの長単位（L3, L2, L1）と係り先文節の先頭から 3 つの長単位（R1, R2, R3）を抽出した．また，文節境界およびその前後の長単位間のポーズ情報として，L3〜R3 のうち隣接した長単位同士のポーズの有無（L3L2, L2L1, L1R1, R1R2, R2R3）を抽出した．なお，抽出される文節には，長単位の数が 3 未満のものもある．その場合，存在していない長単位の情報は欠損値となる．たとえば，係り元文節が 2 つの長単位のみで構成されている場合，長単位 L3 およびポーズ情報 L3L2 は欠損値となる．

アクセント句境界の情報としては，文節をまたぎ 1 つのアクセント句が継続するか否か，およびアクセント句境界の位置情報を抽出した．アクセント句境界の位置の分類としては，長単位 L3〜R3 の内部または末尾，長単位 L3 の先頭，（R1〜R3 末尾以外の）文節末尾，その他とした．

3. 結果

3.1 文節の長単位構成

まず，CSJ から収集された文節の情報について示していく．

係り受け関係にある隣接した文節を CSJ 内で検索した結果，全体で 79,005 箇所あった． 係り元の文節のタイプ頻度は全体で 34,658 例，係り先の文節のタイプ頻度は 39,850 例，両者の文節の組み合わせのタイプ頻度は 69,104 例であった．

次に，係り受け関係にある隣接した文節の長単位の構成をみていく．文節を構成する長単位の数は，各文節によって異なるため，構成要素である長単位の数が 3 未満の文節も含まれる．それぞれの長単位の出現数は表 2 のとおりである．

表 2 にあるように，7 割以上の係り元文節が，2 つ以上の長単位から形成されていることが確認できる．係り元の文節ののうち 3 つ以上の長単位から構成されるものは 15% であるので，実質的に 2 つの長単位から構成されるかかり元文節は 6 割程度となる．係り先の文節は，9 割が 2 つ以上の長単位から形成されているが，3 つ以上の長単位で形成されている文節は 4 割弱であった．したがって，2 つの長単位で構成される文節は 5 割程度となる．

表 2: 文節を構成する長単位の出現数とその割合

長単位	出現数 (%)
L3 以上	12170 (15.40%)
L2	61050 (77.27%)
L1	79005 (100.00%)
R1	79005 (100.00%)
R2	70913 (89.76%)
R3 以上	30052 (38.04%)

3.2 長単位間ポーズの観点から

次に，流暢性の基準の 1 つである文節間・文節内ポーズの観点からの分析をおこなっていく．まず，各長単位間に出現する 0.1 秒以上のポーズの出現数とその割合を表 3 に示す．

表 3: L3〜R3 におけるポーズ（0.1 秒以上）の出現数

pause	L3L2	L2L1	L1R1	R1R2	R2R3
ポーズ無	11791	59059	70918	69218	29289
ポーズ有	379	1991	8087	1695	763
ポーズ率	3.11%	3.26%	10.24%	2.39%	2.54%

この表からわかるように，L3 から R3 の間に出現するポーズは，文節間（L1R1）では約 10%，文節内では約 2〜3% であり，一般的に文節をまたぐと流暢性が低くなることが確認できる．その一方で，文節間のポーズ率が相対的に低い場合，係り元文節と係り先文節の繋ぎ目は一息で発話されやすい，すなわち文節間の流暢性が高いと考えられる．そこで，文節間のポーズ率が文節内平均ポーズ率以下（約 2〜3%）となるような L1R1 の組み合わせを抽出した．収集した 79,005 箇所のうち，L1R1 の組み合わせを語彙素のレベルでみると，27,783 例のタイプがあった．その中で，頻度が 10 例以上であり，かつ文節間のポーズ率が 3% 以下となるような例を収集した．その結果，全体で 38 例の表現を収集することができた（表 4）．

表 4 で確認できるように，流暢性が高く，かつ頻度が 100 以上の語彙素の組み合わせとして，「という事」「の方」「の中」「という風」「って言う」「た時」「言う事」「た場合」「の場合」「の時」「言う風」がある．このうち，「という事」「という風」「って言う」「言う事」「言う風」は，「{ と/って } 言う」形式が用いられており，「機能的な意味を担うような言語表現」として土屋ほか (2014) で挙げたものと共通している．頻度が 100 未満の表現でも，共通した表現として「ている事」「という意味」が挙げられる．他にも，「の研究」「の表」「此の図」「た結果」「の影響」など，学術的な領域で用いられるような表現が見られた．

それ以外に一定の機能性を有するものとしては，「の中」「た時」「た場合」「の場合」「の内」のように命題の条件を指定する表現，「の時」「の方」「の頃」「の日」のように特定の時空間

的領域を指す表現，「ている事」「てる事」「為る事」のように特定の事態や行為を動名詞化する表現が挙げられる．

表 4: 流暢性の高い文節境界

L1R1 語彙素	生起頻度	ポーズ頻度	ポーズ率	L1R1 語彙素	生起頻度	ポーズ頻度	ポーズ率
という/事	1680	47	2.80	の/研究	46	1	2.17
の/方	811	14	1.73	に/近い	45	1	2.22
の/中	406	6	1.48	を/含む	40	1	2.50
という/風	405	4	0.99	てる/事	40	1	2.50
って/言う	401	8	2.00	という/意味	40	1	2.50
た/時	311	8	2.57	此の/図	40	1	2.50
言う/事	168	1	0.60	の/表	38	1	2.63
た/場合	163	2	1.23	其の/人	38	1	2.63
の/場合	146	2	1.37	た/結果	38	1	2.63
の/時	142	4	2.82	から/見る	38	1	2.63
言う/風	127	1	0.79	其の/子	37	1	2.70
もう/一つ	94	1	1.06	も/言う	37	1	2.70
の/内	77	1	1.30	為る/事	37	1	2.70
ている/事	74	2	2.70	を/買う	36	1	2.78
の/頃	72	1	1.39	の/数	36	1	2.78
を/受ける	70	1	1.43	の/影響	36	1	2.78
もう/少し	58	1	1.72	を/与える	35	1	2.86
を/掛ける	55	1	1.82	だ/動作継続	34	1	2.94
の/日	47	1	2.13	に/使う	34	1	2.94

3.3　アクセント句境界の観点から

次に，流暢性のもう 1 つの基準として設けたアクセント句境界情報の観点から分析をおこなっていく．先ほど収集した文節の中で，アクセント句境界の出現箇所を検索したところ，全体で 159,671 箇所あった．これらのアクセント句境界から，言いよどみにともなう語断片やフィラー表現，母音不確定の表現など非流暢性にかかわるアクセント句を除外した結果，154,451 箇所をアクセント句境界として抽出できた．それぞれの出現位置の分布は，表 5 のとおりである．

表 5: アクセント句境界とその句末境界音調の分布

出現位置	L3 先頭	L3 内部	L3 末尾	L2 内部	L2 末尾	L1 内部	L1 末尾	
頻度	340	1118	1666	1943	1341	2177	57966	
出現率 (%)	0.22	0.72	1.08	1.26	0.87	1.41	37.53	
出現位置	R1 内部	R1 末尾	R2 内部	R2 末尾	R3 内部	R3 末尾	文節末尾	その他
頻度	1910	9070	4616	39290	1380	15632	13552	2450
出現率 (%)	1.24	5.87	2.99	25.44	0.89	10.12	8.77	1.59

表 5 を見ると，全体の 4 割近くが L1 末尾，すなわち係り元文節末尾でアクセント句が終了

していることが分かる．その次に出現の割合が高いのは，R2 末尾で 25% となっており，係り先文節では 2 番目の長単位の末尾にアクセント句境界が存在していることが確認できる．

次に，収集したアクセント句境界情報から，係り受け関係にある 2 つの隣接した文節をまたいでいるアクセント句を抽出したところ， 全部で 249 表現を収集することができた．非流暢性にかかわるアクセント句を取り除いたものは，合計で 205 例であった．この 205 例のアクセント句の L1R1 の語彙素のパターンを集計し，頻度 2 以上のものを抽出した結果，表 6 のとおりとなった．

表 6: 文節をまたぐアクセント句における L1R1 の頻度（頻度 2 以上）

表現	頻度	表現	頻度
という/事	69	を/見る	3
と/思う	6	って/言う	2
という/風	5	という/感じ	2
に/成る	5	という/状況	2
何々為る/事	5	に/為る	2
が/有る	3	の/時	2
た/場合	3	の/中	2
ない/成る	3	の/方	2
は/無い	3	考慮為る/事	2

表 6 を見ると，最も生起頻度が高いのは「という事」で 69 例，次に「と思う」「という風」であった．これは，土屋ほか (2014) で挙げられた例だけでなく，文節間のポーズの観点から流暢性が高い表現と一致している．この他にも，2 回しか生起していないが，「の時」「の中」「の方」「という感じ」「という状況」など，前節の分析で挙げた表現と共通した意味的特徴をもつ表現が確認できた．具体的には，命題の条件を指定する表現（「た場合」），特定の時空間的領域を指す表現（「の時」「の方」），特定の事態や行為を動名詞化する表現（「何々為る事」「考慮為る事」）が挙げられる．

次に，文節をまたいだ 205 例のアクセント句境界の位置を見ていく．表 7 にあるように，文節をまたいだアクセント句の 4 割近くが R1 末尾までに境界が存在し，9 割近くが R2 末尾までにアクセント句が終了している．これら 205 例のアクセント句が存在する係り先文節のうち，長単位が 3 つ以上存在するものの数を確認したところ，全体で 140 例あった．これは 205 例の文節の 68.29% を占める．したがって，アクセント句境界の位置は，R3 に出現しにくく，全体（表 5 参照）と比較すると，R1 末尾に終了する傾向が見られる．

表 7: 文節をまたぐアクセント句のアクセント句境界の出現位置の分布

	R1 内部	R1 末尾	R2 内部	R2 末尾	R3 内部	R3 末尾	その他	合計
頻度	22	60	68	30	11	9	5	205
出現率	10.73	29.27	33.17	14.63	5.37	4.39	2.44	100.00

4. 考察

以上の分析の結果，文節をまたぎながらも，一定のまとまりをなしていると考えられる表現として，「{ と/って } いう事」をはじめとする形式的な語彙を用いた表現に加え，「の場合」などの命題の条件を示す表現，「の時」「の方」など特定の時空間的領域を示す表現，「為る事」など事態や行為を動名詞化する表現が確認できた．これらの表現は，2 つ以上の語が文節の境界を越えて韻律的に結合していることから，文節をつなぐ機能的な意味を担っている可能性が考えられる．それに対して，土屋ほか (2014) で挙げられた「ように」「ような」を用いた表現や，思考や伝達にかかわる語彙は，いくつか観察されたものの（「という感じ」）全体的にはあまり観察されなかった．

文節をまたいだアクセント句の句末境界は，R1 末尾と R2 末尾に終了し，R3 で終了しにくい傾向が確認できた．この点からも，文節をまたいで結合している表現は，文節境界に依存的な存在で，文節同士の関係性をあらわす機能的な意味を持つ可能性が考えられる．しかし，今回抽出した表現が実質的にそのような機能を持つか否かを知るためには，それぞれの表現の意味を，全体の文脈を踏まえてより精査する必要がある．これについては今後の課題としたい．

本研究は，これまでに対象としてきた言語表現を大幅に広げての分析となったが，結果的に土屋ほか (2014) での分析結果を再確認するようなかたちとなった．流暢性の観点から文節をまたいで結合している表現を抽出することはできたものの，意味的な機能の分析が十分ではないため，注連ほか (2007) や国立国語研究所 (2001) で挙げられている機能表現と一致していない点も多い．今後は，意味的な分析を進めつつ，先行研究で挙げられている機能表現の発話状況を流暢性の観点から確認していくことで，機能表現の形態的・意味的特徴を探っていくことが課題となる．

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

Fillmore, Charles J (1979). “On fluency.” Daniel Kempler, and William S. Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior*. New York: Academic Press. pp. 85–101.

小磯花絵・伝康晴・前川喜久雄 (2012).「『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.

国立国語研究所 (2001).『現代語複合辞用例集』.

Pawley, Andrew, and Frances Hodgetts Syder (1983). “Two puzzles for linguistic theory: Nativelike selection and nativelike fluency.” *Language and communication*, 191, p. 225.

注連隆夫・土屋雅稔・松吉俊・宇津呂武仁・佐藤理史 (2007).「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』, 14:5, pp. 167–197.

土屋智行・伝康晴・小磯花絵 (2014).「発話の流暢性を踏まえた機能表現の抽出と分析」　『言語処理学会第 20 回年次大会論文集』　pp. 19–22.

土屋雅稔・注連隆夫・松吉俊・宇津呂武仁・佐藤理史・中川聖一 (2007).「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.

関連 URL

「会話コーパス」ホームページ：`http://www.jdri.org/kaiwa/`