ディスコースからみた文末表現抽出

ホドシチェク ボル[1]　阿辺川 武[2]　仁科 喜久子[3]　ベケシュ アンドレイ[4]
[1] 大阪大学　[2] 東京大学
[3] 東京工業大学　[4] リュブリャナ大学
hodoscek.bor.hmt@osaka-u.ac.jp　abekawa@p.u-tokyo.ac.jp
knishina@m06.itscom.net　andrej.bekes@ff.uni-lj.si

概要

本稿では日本語非母語話者の論文作成支援のために学術論文の構造，特に接続表現と文末表現がディスコースマーカーとして重要な役割を果たすことに着目する．まず，学術論文 14 件に対して文末表現と思われる文字列を人手でアノテーションし，接続表現，文末表現と命題との境界に関して定性分析を行った．次に 40 項目の文末表現を 10 機能別に分類し，科学技術系と人文社会系の 6,371 論文に対しパターンを定義して抽出を行い，頻度情報から論文中の文末表現分析の妥当性を検証する．

1　はじめに

本稿は，日本語非母語話者の論文作成支援のために，必要な学術論文の談話構造を語用論の視点で分析する．

語用論の立ち場から，Fraser[1] は「文の意味は、命題的内容と語用論マーカーのセットという二つの部分から構成される」とし，基本メッセージである命題がマーカーによってディスコース内の関係を示すとしている．

(1)　A: Marsha is away for the weekend.
B: ***So***, she won't be available Saturday. [1, p. 188]

(2)　Jane is here. ***However***, she isn't going to stay. [1, p. 187]

(1) は発話者 A の発話から B は，“So”（そうだとすると，だから）で受けて，土曜日の予定は不可能だと推論している．これは日本語の接続表現に対応している．(2) は前文に対して，日本語では逆接「しかし」の意味から始まる文が続く．Fraser はこれらの例の他に，語用マーカー（Pragmatic Marker）の例を挙げている．

(3)　John saw Mary, ***didn't he?***　[1, p. 177]

文末に “didn't he?”を付加することで確認のマーカーとなり，日本語の文末表現の終助詞「ね」に相当する．

(4)　***One hears that*** the jury for the O. J. trial had many internal problems.　[1, p. 183]

“One hears that”は，後に続く命題が伝聞によることを示すマーカーとなっている．

このように Fraser の説く「語用論マーカー」は本稿における接続表現と文末表現の両方に相当するものとなっている．Sarda 他 [2] も同じく語用論の立場で命題の意味内容には関与せずに，文章中の文と文の関係を示す機能を副詞句群（adverbials）とし，接続詞，名詞などを含む複合語となることもあるとしている．

談話分析はさまざまな談話単位（文の要素から意味的パラグラフ「段落」まで）が談話にどのように貢献しているか，または談話をどのように構成しているか，さらにはそのために，さまざまな談話メカニズム（語用マーカーなどの談話標識，指示，遷移など）がどのように働いているかを追求する作業であるが，本稿では文章の中の学術論文という談話ジャンルに属する文章の分析を行う．

2　用語の概説と問題点の再考

阿辺川他 [3] では J-STAGE の学術論文データに見られる文頭にある 5 形態素以内の語句から抽出精査した文字列 523 項目を接続表現として提示し，計量的な分析を行った．また，ホドシチェク他 [4] では，同じ学術論文データを用いて，命題の後に続く文字列から文末表現と思われるものを抽出してディスコース・マーカーとしての機能について計量的分析を行った．しかしながら，これらの先行研究では，「文末表現」に関する定義の曖昧さが課題として残った．そのため，「接続表現」「命題」「文末

表現」からなる文とそれに続く文の連続の中で，特に「命題」と「文末表現」の境界を明確に定義する必要があると考えた．本稿では「命題」と「文末表現」の境界を明らかにするために定性分析および定量分析を行い，学術論文という談話ジャンルにおける構造を考察する．

(5) 意志表現も表現データに示す割合からすると比較的少ないと考えられる．しかし，同時に下記の問題も明らかになったといえる．[1)]

(5) の 2 文において，「しかし」は接続表現，破線部 2 か所はそれぞれ命題，下線は文末表現とする．以下，「接続表現」「命題」「文末表現」の各項目について用例を挙げて概説する．

2.1 接続表現

接続表現は文章の中で文と文，あるいは段落と段落の関係を示す機能表現であり，品詞分類における接続詞「そして，しかし，また」などのほかに「例えば，特に，なお」などの副詞，「要するに，この結果から，いずれにせよ」などの連語も含まれる．

2.2 命題

吉田 [5] は「命題（proposition）」は本来，論理学の概念であり，「1 と 1 の和は 2 である．」「1+1=2」のように異なる表現で，同じ意味内容を表すものが複数ある場合，これらをまとめる意味内容が「命題」と呼ばれるとしている．Fraser のいう基本メッセージである．

2.3 文末表現

Fraser が提唱する Discourse Marker は日本語では命題の前後に分かれて付与されると考えられる．前方にある場合は「接続表現」，命題の後にある場合は「文末表現」となり，ともに「命題」に対する書き手の思考や評価などの態度が示される機能表現である．ホドシチェック [4] では，仁田・益岡 [6] で「モダリティ」と言う概念を談話中の話者の態度を表す表現と定義して論を展開している．

そこで，本稿では，「モダリティ」ではなく，「文末表現」というより広い概念を用いる．英語と日本語の文構造の違いから，英文では ‘can, must, have to, may’ などの助動詞が，日本語では文末辞が書き手の態度を表す．さらに「これを省くこととする」「それが最も多いということになる」などの「（こと）とする」「（ということ）になる」は，書き手の態度を表現しているが，モダリティの基本概念からは逸脱していると思われるからである．(5) の文末表現は「（少ない）と考えられる」「（明らかになった）といえる」と命題部分の用言述部の後の複数形態素からなる文末辞である．一方，(6) のような 3 つの複合形態素からなる複雑な文末辞も見られる．

(6) 現実のロケットの打ち上げにおいて構造振動周波数などは（…）様々な打ち上げに供することができる$_1$ようになる$_2$と期待される$_3$．[2)]

3 文末表現の抽出と分析

3.1 文末表現の分類と特徴

表 1 は機能別に分類した文末表現を示す．この機能分類には「否定」「意志・措置」など，[6] のモダリティ分類にはない分類を加えている．日本語の文末表現は，英語のモダリティにはない文法項目（否定）や，書き手の行為などを含んでいると考えたからである．

人文と科学系の 14 論文に対して文末表現をアノテーションした．文末表現を最末尾辞から形態別に項目として認定した結果，40 種の大分類項目とその派生形である細分類を加え，延べ 105 項目となった．これらを機能別で分類すると 10 カテゴリに分けられた．表 1 は大分類 40 項目とその細分類項目中代表的（高頻度）文末表現を機能別にまとめたものである．例えば，機能分類「ある」の例として「過度に励起しないように注意する必要がある」などの用法が多く見られた．機能名は仮の名称も含まれている．

3.2 抽出方法

文末表現リストの正確な抽出を成すため，spaCy[7] の Matcher クラスで複雑なパターンを簡潔に記述できる簡易ドメイン特化言語（DSL）を考案し，Python のコードで各文末表現を定義した．その一例を図 1 に示す．

1) 科学技術論文コーパス内：乾 裕子，村田 真樹，内元 清貴，井佐原 均. 表層表現に着目した自由回答アンケートの意図に基づく自動分類．自然言語処理，Vol. 10, No. 2, pp. 19–42, 2003.

2) 分析資料：澤井 秀次郎, 松田 聖路．動翼を用いた観測ロケットの適応型姿勢制御系の設計とハードウェア試験．宇宙技術, Vol. 1, p. 10, 2002.

表 1　文末表現の機能別分類リスト

機能	大分類項目（細分類）
断定・存在	である（のである, なのである, ものである, ことである）, ある（必要がある, 場合がある）
推量・意志	う, いう（といえる）, 考える（と考えれる）, 思う（と思われる）, うる, 推測する, 予想する, 推定する
意志・措置行為	とする（こととする）, 示す, 意味する, 指摘する, 意図する, 仮定する, おく（とおく）
認識	となる, なる（ことになる）, わかる, あげる（があげられる）, 示唆される, 示される, 多い（ことが多い）, 確認する, 指摘する（が指摘される）, 知る（知られる）, 報告する（報告している）, 理解する（理解されている）
可能（性）	できる（ことができる）, いう（といえる,V という）, かもしれない
疑問	か（のか, であろうか, ないだろうか）
否定（部分否定）	ない（ことはない, ものではない, わけではない）, 限らない（とは限らない）
願望・期待	たい, 望ましい, 期待する（ことが期待される）
容認	て（も）よい
仮の処置	仮定する, とおく

```
VERB_PHRASE = OR(  # 動詞（句）
    AND(
        m(TAG=m(IN=AND("動詞-非自立可能", "動詞-一般"))),
        m(POS="AUX", OP="*",
          NORM=m(NOT_IN=("無い"))),
    ),
    AND(
        m(TAG="名詞-普通名詞-サ変可能"),
        OR(
            m(POS="AUX", OP="+",
              NORM=m(NOT_IN=("出来る"))),
            SURU,
        ),
    ),
    AND(m(NORM="為る"), m(POS="AUX", OP="*")),
)

V_TO_SURU = AND(  # 意志・措置行為の文末表現「とする」系
    VERB_PHRASE,
    m(ORTH="と"),
    m(NORM="為る"),
)
```

図 1　形態素に基づく文末表現の「V とする」定義例

この DSL を使用することで，Matcher が期待するデータ形式に自動的に展開される入れ子構造の選択肢を簡潔に記述することが可能で，spaCy の API における制限を克服するものである．例えば，上記の図 1 で「V とする」表現の派生型に一致させるためには，動詞だけでなく，サ変名詞及びそれらの助動詞も考慮に入れる必要があり，このように動詞句が他の文末表現の定義で再利用可能となる[3]．

抽出するにあたり，文末表現のパターンが文末以外に抽出された場合の処置法として，文の中の出現位置が 0.7 以上の表現のみを文末表現候補とする．接続表現の抽出は文頭に現れる[4]．GiNZA[8] の ja_ginza モデル（v5.1）を用い，論文の各文から文末表現を抽出した．

3.3　抽出結果

表 2　ジャンルごとの文末表現とその機能・細分類および百万文あたりの調整頻度（人社=人文社会学論文は文数 447,645 で，科技=科学技術論文は 1,182,181 文数で調整）*

機能	細分類	人社	科技
断定・存在	である	130,773	46,988
	ある	15,561	9,199
推量・意志	う	24,832	2,064
	いう	18,376	3,338
	考える	16,949	12,722
	思う	7,477	1,951
	うる	2,410	546
	推測する	172	80
	予想する	71	40
	推定する	18	11
意志・措置行為	とする	41,705	21,235
	示す	1,970	1,560
	意味する	1,398	507
	指摘する	1,206	34
	意図する	85	22
認識	となる	19,985	9,340
	なる	8,576	1,724
	わかる	5,464	6,052
	あげる	2,174	1,068
	示す	1,430	640
	示唆する	1,360	453
	多い	1,083	1,048
	確認する	983	821
	指摘する	398	53
	知る	150	221
	報告する	76	103
	理解する	4	0
可能（性）	できる	23,199	18,950
	いう	6,237	1,079
	かもしれない	1,276	169
疑問	か	13,937	1,051
否定（部分否定）	ない	13,124	2,778
	限らない	386	363
願望・期待	たい	7,953	1,650
	望ましい	297	431
	期待する	201	661
容認	てよい	326	140
仮の措置	仮定する	176	532
	とおく	9	25

*各細分類の素頻度において χ^2 検定をボンフェローニ法で補正した有意差を有意水準0.001, 0.05, 0.1 でそれぞれ下線，波線及び破線で調整頻度の多いジャンルを示している．

表 2 は，3.2 で述べた用法によって機械的に抽出した結果である．人文社会学論文（以降，人社）と

3）　全パターン及び抽出の詳細は https://github.com/borh/dm-annotations で公開している．

4）　接続表現の抽出の詳細に関しては阿辺川他 [3] を参照

科学技術論文（以降，科技）を比べると全体として「人社」論文における文末表現の使用頻度が高いことがわかる．「科技」では客観的な事実や実験結果をもとに断定的に論を進めるのに対し，「人社」では調査や観察を通して対象の内面を間接的に推測する手法が多く行われている．そのため「人社」では結果に対し著者の解釈の幅が広く，多様な文末表現の表出につながったのではないかと推測される．

次に「人社」と「科技」の論文の間で大きく使用頻度が異なる機能分類に着目し，その原因を探る．機能分類「疑問」は「人社」の使用頻度が「科技」の 10 倍以上もある．「人社」論文では社会文化に対する問題提起から論を進める構成や，考察の中で明確な根拠がなく筆者による推測が入り込むことにより「疑問」の文末表現が多用されるのでないかと考えられる．

一方，機能分類「仮の措置」では「科技」論文での利用頻度が高い．これは実験や証明の事前準備としてこの文末表現を使用する機会が多く，「科技」論文特有の表現と言える．

3.4 考察

「科技」の論文は，現象から仮説によって事柄の実相を想定し，実験によって証明し，将来実現することを期待するという記述内容が伺える．一方，「人社」の論文は，社会文化の現象を解釈し，将来に向けて改良する方策を示すという様子がうかがえる．

4 節のパイロットスタディで観察した論文 1 の緒論と結論の段落の中で，命題を挟む接続表現と文末表現が連続して出現する様相を示す．論文中の同様の例を定量的に分析することで，段落全体のパターンの定型化を掴む可能性が考えられる．緒論・本論・結論の各部のパターン，さらにはジャンル別の論文のパターンを記述することである．

その方法として文章中の接続表現，文末表現の出現の流れをまとめていく [9] の分析法のほかに，MOVE 分析 [10, 11] を援用することも有効だと思われる．

4 接続・文末表現連鎖パターンの定型化

Bekeš ら [9] の研究では，文章中の接続表現の遷移を辿り，展開の構造を示した．今後は，接続表現と文末表現を組み合わせることで，より詳細な構造の提示が可能であると考える．網羅的な文末表現の抽出ができれば，様々な視点からの談話分析が可能となる．例えば，論文中の緒言と結論における文の展開を記述することも可能となる．ここで，緒言と結論における文の展開の分析の例を提示する．

分析論文 1[a]

a 立本 英機, 成田 高秀, 相川 正美. 感潮河川域における底質中の形態別リンの分布特性. 日本化学会誌 (化学と工業化学), Vol. 2002, No. 3, 2002.

緒言 s_1 近年 (命題) 取られている. s_2 中でも (命題) ものの, (命題) 懸念される. s_3 そのために (命題) 重要である. s_4 また (命題) 少ない. s_5 そこで (命題) 調査した. s_6 (花見川は…整っている) と思われる. s_7 また (命題) 加えた. s_8 一方 (命題) 抽出される. s_9 (その生成機構は…) 検討されている. s_{10} それら (命題) と予測される. s_{11} しかし (命題) 少ない. s_{12} そこで (命題) 試みた. (中略)

結論 s_{116} (花見川流域における…再現を) 試みた. s_{117} その結果, 次のようなことが明らかになった. → (箇条書き 4 項目)

記号の説明：s_i は文 i, 接続表現, 文末表現, それ以外文末表現の候補や談話の展開に関わるもの

緒言では，研究課題の所在と理由，解決のために方策，解決の見通しを述べる機能語「そのために，少ない，思われる，予測される，しかし，そこで」が展開されている．結論では，「結果，明らかになった」という機能語で，調査結果を明示している．

このような形で学術論文から接続表現と文末表現の連鎖パターンを抽出し，高頻度で出現するパターンを定型化できれば，論文の構造理解や執筆の参考になると考えられる．

5 終わりに

学術論文作成支援を目標として，論文構造を語用論の視点から接続表現と文末表現に着目し，分析を行った．その中で，本稿では学術論文作成上で重要な文末表現のリスト化に注目し，命題を挟む接続表現と文末表現の連なりの中で，高頻度で働きが重要と思われる文末表現 105 項目を提示し，さらに科学技術論文と人文社会系論文の差異も示した．文章中でこの組み合わせがどのように展開するかを観察し，各分野の学術論文作成に有益な情報を提示するかを明らかにするところまでは至らなかったが，今後，研究を深めていくこととする．

謝辞

本研究は JSPS 科研費 JP23K00629 の助成を受けたものである．ここに謝意を表す．

参考文献

[1] Bruce Fraser. Pragmatic markers. **Pragmatics**, Vol. 6, No. 2, pp. 167–190, 1996.

[2] Laure Sarda, Shirley Carter-Thomas, Benjamin Fagard, and Michel Charolles. Adverbials: From predicative to discourse functions. In **Adverbials in Use: From Predicative to Discourse Functions**, Corpora and Language in Use, pp. 17–34. Presses universitaires de Louvain, 2014.

[3] 阿辺川武, 仁科喜久子, 八木豊, ホドシチェック・ボル. 日本語接続表現の計量的分析に基づく指導法の提案. 計量国語学, Vol. 32, No. 7, pp. 387–402, 2020.

[4] ホドシチェク・ボル, 阿辺川武, 仁科喜久子, ベケシュ・アンドレイ. 学術論文形成を支える接続表現と前後文末モダリティとの共起構造―談話分析の視点から―. 計量国語学, Vol. 34, No. 1, pp. 1–16, 2023.

[5] 吉田夏彦. 命題・意味の項目. 国語学大辞典, 第 327 巻. 1989.

[6] 仁田義雄, 益岡隆志. 現代日本語文のモダリティの体系と構造. 仁田義雄, 益岡隆志（編）, 日本語のモダリティ, pp. 1–56. くろしお出版, 1989.

[7] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[8] 松田寛. GiNZA - Universal Dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.

[9] Andrej Bekeš, Bor Hodošček, Kikuko Nishina, and Takeshi Abekawa. Distant co-occurrence patterns of connectives: a corpus study of formulaicity in Japanese. **Acta Linguistica Asiatica**, Vol. 13, No. 2, pp. 9–38, Jul 2023.

[10] John M. Swales. **Genre Analysis**. Cambridge Applied Linguistics. Cambridge University Press, 1990.

[11] John M. Swales. **Research Genres: Explorations and Applications**. Cambridge Applied Linguistics. Cambridge University Press, 2004.