比較文の意味解析のための「深い」係り受け関係の解析

窪田悠介[1]　林則序[2]　天本貴之[3]　峯島宏次[3]
[1] 国立国語研究所　[2] 東京大学　[3] 慶應義塾大学
kubota@ninjal.ac.jp　hayashi-lin@g.ecc.u-tokyo.ac.jp
amamoto@keio.jp　minesima@abelard.flet.keio.ac.jp

概要

言語学的に複雑な現象に関する意味解析を行うための方法として、「深い係り受け」という概念を提案し、その有効性を検証したパイロット研究の結果を報告する。「深い係り受け」とは、一言で言って抽象的な意味関係に関わる情報であり、理論言語学では統語変換などの複雑な操作によって規定される。このような情報を、深層学習などの最近の機械学習の手法でどの程度正確に解析できるかを検証した研究は未だ存在しない。本研究では、日本語比較文の分析に関わる「深い係り受け」情報の判定器として、深層学習モデルと (言語学的知識に基づく) 規則ベースモデルの二種を実装し、その比較を行った。

1　はじめに

深層学習の手法や大規模言語モデルが手軽に利用可能になったことで、自然言語処理 (NLP) 技術を援用して言語理論の問題に取り組む研究が活発化している。しかしながら現在までの研究は、既存の NLP 技術で解ける形に言語理論の問題を規定し直す形のものが多い。このため、「統語変換の概念を実装した頑健なパーザは構築可能か？」といった、理論言語学側の核心的問題を立脚点とした研究は、一部の先駆的試み [1] を除いて端緒にすらついていない。

本研究では、理論言語学研究に動機づけられた概念として、「**深い係り受け**」という概念を提案する。統語解析から意味解析へのパイプラインの途中に深い係り受けの解析レイヤを組み込むことで、複雑な意味解析を機械に実装可能な形で行うための見通しが立つ。具体的には、深い係り受けの解析自体が、多くの場合、既存の機械学習の手法を援用することで比較的容易に高精度で可能であり、また、句構造解析や含意関係認識などの、この解析レイヤの上下に接合する他のコンポーネントに関しても既存の高精度な解析器がそのまま利用可能となる。本論文では、このような設計の意味解析システムのプロトタイプとして構築した、日本語の比較文の意味解析のための深い係り受けの解析器を報告する。

2　理論的前提

比較文の意味を正確に解析することは、自然言語の意味論における重要な課題である [2, 3]。また、比較文は (1) に見られるような含意関係をもたらす言語表現であるため、NLP の含意関係認識においても重要な課題の一つである [4]。

(1)　P1: 太郎は次郎より背が高い。
　　P2: 次郎は三郎より背が高い。
　　C: 太郎は三郎より背が高い。

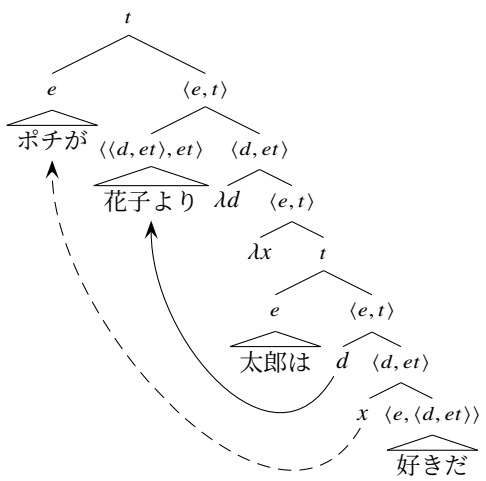
比較文の意味解析には、表面的な係り受け関係だけでなく、以下三種類の情報の判定が必要となる [5]。

(2)　a. 何と何が比較されているか
　　b. 比較の尺度 (スケール) は何か
　　c. 差分がいくらか

以下、便宜的にこの情報を「**深い係り受け**」と呼ぶ。

理論言語学では、比較文を (3) に示す「**寄生スコープ (parasitic scope)**」[6] という複雑な統語変換操作で分析する方法が提案されている [7, 8]。

(3)



(3) の分析図は一見複雑に見えるが、実際には、既存の句構造解析器の出力と (2a–c) に示した深い係り受けの情報があれば、単純な構文木の変換により一意に復元可能である。つまり「寄生スコープ」とは、深い係り受けと浅い係り受けをどう組み合わせれば文の意味が得られるか、ということに関する母語話者の直観的知識を言語学者が明示的に規則化したものにほかならない。この手法により、比較文のような複雑な意味現象に関して、比較的単純な方法で高階論理の正確な意味表示を導くことが可能となる。

具体的には、(4) の文に関しては、(2a–c) の情報があれば、単純な構文木の変換により、それぞれ (5) の論理式を得ることができる ((4) で用いたラベルについては 3.1 節を参照)。

(4) a. 太郎は花子よりprej ポチcont が好きだdeg。
b. 警察官cont の初任給は他の公務員よりprej 少しdiff 高いdeg。

(5) a. **max** d[好き(花子)(太郎)(d)] > **max** d[好き(ポチ)(太郎)(d)]
b. **max** d[多い(初任給(警察官))(d)] > **max** d[多い(初任給(他(公務員)))(d)]

文法理論に基づく高階論理解析器 ccg2lambda [9] などで採用されている既存の CCG 構文解析器をそのまま用いて、(3) の分析図や対応する (5) の高階論理式を得る方法は自明ではない。寄生スコープは、比較文だけでなく、(「同じ/別の」「それぞれ」「平均して」「合計で」などの) 複数名詞句や等位構造と連動して複雑な意味解釈をもたらす一連の言語表現の分析にも有効であることが知られている [10]。従って、寄生スコープ分析に基づく自動解析の方法を模索することは、理論言語学研究への NLP 技術の援用のための道筋をつける目的のみならず、NLP の推論タスクなどへの理論言語学的知見の活用のためにも一定の意味があると考えられる。

3 深い係り受けの判定器の比較

寄生スコープ分析に基づく比較文の自動解析器を作り、それを用いてコーパスなどに現れる任意の文を解析するためには、以下の二つの情報の両方を正確に推定することが必要となる。

- **浅い係り受け:** 表面的な依存関係/句構造
- **深い係り受け:** (2) の比較文の意味に関わる情報

この二つのうち、浅い係り受けに関しては既存の句構造文法 (HARUNIWA2 [11] など) や CCG の解析器 (depccg [12] など) を用いることができる。

深い係り受けは (6) の特徴を持つ情報である。

(6) a. 抽象的な意味関係に関わる情報である
b. 部分的に文法的マーキングで表示される

深い係り受けの推定は、(6a) の特徴のため本質的に困難なタスクだが、(6b) の特徴を持つことから、形態・統語的特徴や意味的類似度などの情報により、ある程度の精度で推定できることが予想される。このため、言語学的知見と NLP 的手法を組み合わせたアプローチが特に有効と考えられる。しかしながら、我々の知る限り、深い係り受けの推定 (や類似タスク) に関して、異なる特徴を持つ複数の手法を比較し、それらの性能や利点・欠点を具体的に検討したした研究はいまだ存在しない。

本研究では、深い係り受けの推定に関して、言語学的知見に基づいて規則を手で書いた規則ベースの判定器を用いる手法と、アノテーション・データを用いた深層学習のラベル認識タスクとして解く手法を構築し、両者の解析結果を比較した。以下、タスクの定義と、それぞれの判定器の詳細を記述する[1)]。

3.1 タスクの定義

2 つのモデルにおけるトークン化の相違や、係り受け関係の利用可能性の相違があるため、タスクは文字列の区間に対するラベル付けタスクとして定義する。具体的には、日本語の文である任意の文字列に対し、正解データにおいて付与されているラベルと判定器が予測したラベルの一致率を計算する。

(7) **ラベルの種類**
prej: 比較句 (例: 「花子より」)
cont: 比較句と対になる句 (例: 「ポチが」)
diff: 差分表現 (例: 「少し」「3cm」)
deg: 程度述語 (例: 「好きだ」「背が高い」)

(8) **正解データの例**
妻が仕事に精出す一方、[赤沼は]cont [それより]prej [もっと]diff [忙しい]deg 。
(BCCWJ LBl9_00238, 5950)

1) ソースコードおよびモデルについての情報は https://github.com/ABCTreebank/comparative-ner-utils/releases/tag/NLP2023 にて公開中である。

3.2 規則ベースの判定器

既存の一般的な言語処理ツール（形態素解析、依存構造解析）を使用して規則ベースの判定器を実装した。まず、入力文を GiNZA [13] によって形態素解析し、依存構造解析を行う。この依存構造は日本語 Universal Dependencies (UD) [14] に基づくものであり、例えば、(4b) の例には付録図 1 に示したような出力（依存構造木）が得られる。

これを用いて、次の順序で特定の条件を満たす構成素に素性を割り振る。構成素の意味的な類似性については、GiNZA が提供する単語埋め込みによるコサイン類似度を利用した。

1. **prej**: 依存構造木から「N より」「N {と・に} 比べ」における名詞 N を特定し、それを主要部とする部分木を prej とする。(4b) では「公務員」が N であり、「他の公務員より」が prej となる。
2. **deg**: N の親（「高い」）を deg とする。
3. **diff**: deg の子に副詞（「少し」）があれば、それを diff とする。
4. **cont**:「X の方」という句がある場合は、X を主要部とする部分木を cont とする。「X の方」がない場合は、prej、deg、diff に含まれていない要素の中で、N ともっとも類似度が高い要素を X とする。その上で、X が名詞であれば、X を主要部とする部分木を cont とし、X が名詞でない場合、cont は空とする。(4b) の場合、「警察官」が cont となる。

3.3 機械学習に基づく判定器

学習データ　BCCWJ [15] から、「より」と「{と・に} 比べ」を含む文で、比較構文に該当する可能性があるものを 3,460 文抽出した。付録表 1 に内訳を示したとおり、これらの文は、複文 (連用節、連体節) をなすものを含み、また、比較文でないデータも含んでいるため、3.1 節で定義した深い係り受け判定のタスクは一定の複雑さをもつものであると考えられる。これらの文に対して、アノテーションを上の (8) に示したブラケット形式で、手動で施した。アノテーションデータのうち、評価用に 350 文（全体のおよそ 10%）をランダムに選出し、残りの 3,110 文を学習に用いた。

モデル　BERT モデル [16] [2] の上に、トークンを分類する線形ニューラルネットワークを加えたものをモデルとした。これは、固有表現抽出（Named Entity Recognition, NER）と同じモデルの構成である。単語埋め込みベクトルの次元数は 768、入力長を 256 に設定した。分類ラベルは、空ラベル（O）および (7) に示したラベル（prej, cont, diff, deg）の 5 つで構成した。

モデルの学習　上記学習データを用いて学習を行った。学習時のエポック数は 27、ミニバッチサイズは 16、学習率は 5.0×10^{-5} とした。

4 結果・考察

4.1 評価方法

評価用 350 文について、規則ベースモデルと機械学習モデルの 2 種類で予測をし、正解データと照合した。予測された素性のスパンと正解データのスパン同士の可能なマッチングパターンを網羅し、個々のマッチングペアに、下記の判定に自然に沿うようなコスト付けを行い、最小コストのマッチングを、線形割当問題ソルバー[3] を用いて算出した。マッチングペアの判定は次の 4 種類からなる：

- CORRECT（完全一致）
- SPURIOUS（余分な予測）
- MISSING（正解スパンの予測の失敗）
- WRONG_SPAN（スパンが重なるが不一致）

結果の集計のために、MUC-5 評価メトリクス [17] のうち、再現率・精度メトリクス（recall-precision metrics）を用いた。メトリクスには strict と partial の 2 種類を設け、strict における正解は CORRECT のみ、partial における正解は CORRECT+0.5WRONG_SPAN とした[4]。評価結果は付録表 2 および表 3 の通り。

4.2 モデル間の比較

全般的に機械学習モデルのほうがスコアが高いが、素性ごとに 2 つのモデルの差が大きく異なる。

prej はどちらのモデルも F1 値が 80 以上になった。prej は「より」「{と・に}比べ」を手がかりとして特定できる。機械学習モデルはこの点の学習に成功している。一方、規則ベースモデルでは、再現率は機械学習モデルを超えるが、精度が (strict、partial どち

2) 日本語 BERT モデルは https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking を用いた。

3) scipy.optimize.linear_sum_assignment を使用した。

4) [18] の partial boundary matching と違い、素性が不一致で範囲が一致するスパンのペアは正解として扱わない。一般的な固有名認識と違い、我々にとっては素性の不一致は重大な間違いだからである。

らの基準でも) 機械学習モデルに比べ 10 ポイント以上落ちている。その主な原因は、参照している形態素解析、依存構造解析の品詞情報が比較のヨリと起点のヨリ（例：「ロンドンより最近到着した」）を区別していないことにある。起点のヨリに関する過剰な予測を除外すると、精度 (strict) は素性の単純平均で 13.1 ポイント上昇する[5)]。

deg に関して規則ベースモデルの再現率が機械学習モデルに対して劣る主な要因としては、ヨリ句と deg の間に他の述語が介在している場合に典型的に見られる解析エラーが挙げられる（例：「[中央区に比べ]prej、西区のはずれに位置する平和は空気が[澄ん]deg でいる[6)]」で「位置（する）」を deg と予測する）[7)]。規則ベースモデルの diff に関するエラーの要因としては、prej と同様、品詞情報の不足により、差分表現を特定することが難しい点が挙げられる。また、規則ベースモデルは、3.2 節で述べたように prej から順に素性を判定するため、上述の prej 誤判定が他の素性の誤判定（精度の低下）の一因になっている。これらの点はすべて、規則ベースの手法の古典的な問題の一種である。本研究が対象とする、一見単純な言語学的知識で解けそうなタスクでも、アノテーション・データを用いた機械学習が、タスクに特化した教師データなしの規則ベースの手法と比べて実データに対して明らかに頑健であることを示しており興味深い。

一方で、構造を明示的に参照していないせいで機械学習モデルが判定に失敗している例もある。たとえば、付録 (9a) では cont はブラケットで囲った名詞節全体だが、モデルは名詞節内の非連続な区間を cont と予測する。また、(9b) の「黒人」は従属節の主語であり、比較構文をなす主節の要素ですらないが、誤って主節述語に対する cont と予測されている。

4.3 素性ごとの比較

4 つの素性のうち、判定が最も困難で、かつ意味解析にとって重要なものは、統語的・形態的な手がかりが乏しい cont である。これに関しては、2 つのモデル双方で、予想通り 4 つの素性の中で F1 が最も低くなった。この主な要因としては、次の 3 つがある（具体的な例文は付録 (10) を参照）。

意味的類似度の過剰重視 (10a) で比較の対象は時点だが、言語表現として明示されないので、「cont なし」が正解である。しかし、prej 内の国名「フランス」につられて、「ベトナム民主共和国（大統領）」が cont と予測される。（機械学習・規則ベース両方）

cont が名詞句の一部であるケース (10b) においては、比較の正しい対比は「コウヤマキ cont」と「他の五樹種 prej」だが、モデルは「コウヤマキ」を含むより大きな名詞句「…の耐陰性（…）」全体を cont と予測する。（規則ベースのみ）

文脈情報の不足 前文脈、および「おっとりした」から、(10c) の正解は「娘よりも prej」/「花 cont」であると判断できる。しかし、モデルは一般的に cont になりがちな主語名詞句「おっとりした母」を cont と予測する。（機械学習・規則ベース両方）

5 結論

「深い係り受け」の解析を実証的に検証するため、日本語比較文の意味解析を題材に、機械学習と規則ベースの 2 つのモデルを構成し、性能を比較した。以下簡単に結論をまとめ、今後の方針を述べる。

まず、全般的に機械学習モデルのほうが性能が高いが、いくつかの弱点に関しては、異なる種類の言語学的情報を明示的に参照することで性能の改善が見込める。係り受け解析の参照など ((9) 参照)、規則ベースモデルの要素を何らかの形で取り入れるのが有効と考えられる。また、cont の予測の改善のためには、(10c) が示唆するように前文脈を考慮する必要がある。文脈情報の利用で性能向上が見られるかを今後検討する。

「深い係り受け」の解析の先には、これを用いた意味解析が目標としてある。直近の課題は、ccg2lambda などによる、論理式を介した含意関係判定器への接合である。人間の言語処理が、連結した複数のコンポーネント間の制約の最適化問題として規定できることはほぼ疑いの余地がない。深い係り受けのレイヤを立てた意味解析パイプラインは、この設計の機械実装可能な近似と位置づけられる。将来的な課題は、定量的な性能検証が可能な自然言語の意味処理に関するモデルとして、このような設計の機械を構築することである。

5) 対して、機械学習モデルにおいては、4.6 ポイントの上昇にとどまる。
6) BCCWJ LBp9_0011, 1040
7) このほかに日本語 UD とのアノテーション方針の違いの影響もある。例えば、「[さっきより]prej 部屋の中が [明るく]deg なったような気がした」では、日本語 UD の標準的な依存構造では、prej の「さっきより」は、「明るく」ではなく「なっ（た）」を修飾する。

謝辞

本研究は JSPS 科研費 21K00541、国立国語研究所共同研究プロジェクト「計算言語学的手法による理論言語学の実証的な方法論の開拓」、JST CREST JP-MJCR2114 の支援を受けたものである。研究の初期段階において助言を下さった吉川将司氏に感謝する。

参考文献

[1] John Torr, Miloš Stanojević, Mark Steedman, and Shay B. Cohen. Wide-coverage neural A* parsing for Minimalist Grammars. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2486–2505, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Christopher Kennedy. Comparatives, semantics of. In *Encyclopedia of Language and Linguistics*, pp. 690–694. Elsevier, Oxford, 2 edition, 2005.

[3] 澤田治. 比較構文の語用論. 澤田治美（編）, ひつじ意味論講座 2: 構文と意味, pp. 133–155. 2012.

[4] Izumi Haruta, Koji Mineshima, and Daisuke Bekki. Combining event semantics and degree semantics for natural language inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1758–1764, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[5] Omid Bakhshandeh and James Allen. Semantic framework for comparison structures in natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 993–1002, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[6] Chris Barker. Parasitic scope. *Linguistics and Philosophy*, Vol. 30, No. 4, pp. 407–444, 2007.

[7] Christopher Kennedy. Modes of comparison. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, Vol. 43, pp. 141–165, 2009.

[8] Ai Matsui and Yusuke Kubota. Comparatives and contrastiveness: Semantics and pragmatics of Japanese *hoo* comparatives. In *Proceedings of Formal Approaches to Japanese Linguistics 5*, pp. 126–139, Cambridge, MA, 2010. MITWPL.

[9] Koji Mineshima, Ribeka Tanaka, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Building compositional semantics and higher-order inference system for a wide-coverage Japanese CCG parser. In *Proceedings of EMNLP 2016*, pp. 2236–2242, Austin, Texas, 2016. Association for Computational Linguistics.

[10] Yusuke Kubota and Robert Levine. *Type-Logical Syntax*. MIT Press, Cambridge, MA, 2020. Available Open Access at https://direct.mit.edu/books/book/4931/Type-Logical-Syntax.

[11] Butler Alastair Horn, Stephen Wright and Kei Yoshimoto. Keyaki treebank segmentation and part-of-speech labelling. 言語処理学会第 23 回年次大会発表論文集, pp. 414–1148, 2017.

[12] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG parsing with a supertag and dependency factored model. In *Proceedings of ACL 2017*, pp. 277–287, 2017.

[13] 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.

[14] 浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇有吾, 松本裕治. Universal dependencies 日本語コーパス. 自然言語処理, Vol. 26, No. 1, pp. 3–36, 2019.

[15] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[17] Nancy Chinchor and Beth Sundheim. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993.

[18] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 341–350, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

付録

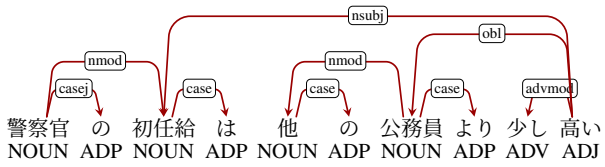


図 1: 依存構造木の例

	連体	連用	その他比較	非比較	合計
より	344	289	1111	778	2522
比べ	55	103	487	293	938
合計	399	392	1598	1071	3460

表 1: アノテーションデータの内訳

				WRONG	strict			partial		
	CORR	SPUR	MISS	_SPAN	精度	再現率	F1	精度	再現率	F1
prej	228	29	8	21	82.0	88.7	85.2	85.8	92.8	89.2
cont	119	53	23	35	57.5	67.2	62.0	65.9	77.1	71.1
deg	219	48	41	10	79.1	81.1	80.1	80.9	83.0	81.9
diff	66	9	7	10	77.6	79.5	78.6	83.5	85.5	84.5

表 2: 機械学習モデルの結果

				WRONG	strict			partial		
	CORR	SPUR	MISS	_SPAN	精度	再現率	F1	精度	再現率	F1
prej	239	86	4	14	70.5	93.0	80.2	72.6	95.7	82.6
cont	50	104	91	36	26.3	28.2	27.2	35.8	38.4	37.1
deg	158	159	100	12	48.0	58.5	52.8	49.8	60.7	54.8
diff	52	62	24	7	43.0	62.7	51.0	45.9	66.9	54.4

表 3: 規則ベースモデルの結果

(9) （機械学習モデルに特有の失敗、下線~~cont~~ が誤った予測）

a. しかし [論理よりも経験よりも何よりも]prej [たいせつ]deg なことは、
[子が~~cont~~ 親から信用されること~~cont~~]cont である。 (BCCWJ LBl1_00024,39960)

b. 黒人~~cont~~ は出生率は増加しているが、[ニューヨークを去る人]cont が [移住してくる人よりも]prej [多い]deg。 (BCCWJ LBl3_00134,8920)

(10) （cont の予測の失敗、下線~~cont~~ が誤った予測）

a. **意味的類似度の過剰重視**
ベトナム民主共和国大統領~~cont~~ が帰りついた祖国の情勢は、[フランスへ向った時よりも]prej [更に]diff [悪化]deg していた。 (BCCWJ LBc2_00023, 22980)

b. **cont が名詞句の一部であるケース**
このことは、[コウヤマキ]cont の耐陰性（少ない光に耐えて育つ能力）が~~cont~~、[他の五樹種にくらべて]prej、[いちじるしく]diff [高い]deg ことを示している。(BCCWJ LBk6_00008, 23250)

c. **文脈情報の不足**
おっとりした母は~~cont~~、[娘よりも] prej[花に]cont [関心]deg がむいている。(BCCWJ LBh0_00004, 74220)