コーパスからの呼応表現自動抽出のための正解データ作成

木田敦子[1]　　山本英子[2]　　神崎享子[2]　　井佐原均[2]

akida@ibs.or.jp　　{eiko, kanzaki, isahara}@crl.go.jp

1 計量計画研究所　　2 通信総合研究所

1　はじめに

文の意味は述語によって決まるという観点から見ると、英語などの SVO 型言語では述語が出現した時点で文内容が把握でき、後続要素の予測がつきやすいと説明できる。これに対して、日本語のような SOV 型言語では、述語が目的語などの後に出現するため、後続要素の予測がつきにくいということになる。だが、日本語でも文を読む時や日常会話において、理解に困難はなく、人は各要素の出現順序に従って、進行していく文を漸進的に理解していると考えるのが自然である。我々は、これを漸進的文理解と呼ぶ。SOV 型言語である日本語において、この漸進的文理解を可能にする要因の一つに、「呼応」という構文構造があると考えられる。

「呼応」とは、「決して行かない」の「決して」と「ない」のように、「文中で前にある特定の語が用いられると、後にそれに応じた特定の語、言い方が現われること」(『国語学大辞典』「呼応」の項より)である。本稿では、以下、「前にある特定の語」を「呼応」の「呼要素」、「後のそれに応じた特定の語、言い方」を「応要素」と呼ぶ。

我々は、呼応関係にある表現を集めて、呼応表現データの作成を試みている(Kida et al. 2003)。これまで、客観的基準と実例に基づいて実用規模の呼応表現データを作成する研究は行われていなかった。そこで本研究では、コーパスから呼応表現を自動抽出することで、客観的な基準に拠る網羅性の高い呼応表現データの作成を試みている。このようにして作成されたデータは、日本語研究のための基礎資料として役立つ。また、対話処理システムに求められる漸進的文理解(木田ほか 2001)のための基礎データとなることも考えられる。

本稿では、自動抽出結果の判定に用いる正解データの作成方法および作成したデータに対する分析結果について報告を行う。呼応表現データの作成にあたり、本研究では直観で気づくことが困難な呼応表現も発見的に抽出することを目指している。そこで、正解データ作成の第一段階では、呼応関係を「文中である特定の語が用いられると、後にそれに応じた特定の表現が表われる関係」と単語－形態論レベルの形式として緩く定義し作業を行う。次の段階で、作成した正解データと自動抽出結果(山本ほか 2004)を比較検討し、直観で発見しにくい表現を追加するなどのブラッシュアップを行う。このように、人手作業と自動抽出の双方の精度を上げていくことで、質の高い網羅的なデータの作成を目指す。

2　陳述副詞の呼応

中世以前の日本語には、係助詞と文末の活用形との形態的な呼応関係である係り結びの用法が見られた。この用法により、係助詞が出現した時点で文末の予測が可能であった。

(大野 1993)は、古語の係り結びに代わって文の行く手を予告する機能を持つ表現として、現代語にも「ある種の副詞」が存在することを示唆している。ここで大野が「ある種の副詞」と述べた副詞は、いわゆる陳述副詞に相当すると考えられる。

陳述副詞は、(山田 1908)によって設けられた品詞分類である。山田は呼応関係を持つことを陳述副詞の分類基準にしている。(橋本 1959)は、山田が陳述副詞に分類した語の中には文末辞がつかない無標形式と共に出現するものがあることを指摘し、述語との呼応が陳述副詞であることの基準にはならないと主張している。橋本の主張に対し(工藤 1982)は、呼応は単語－形態論レベルの形式ではなく、文の中で他の一定の単語と結合しつつ機能している述語－構文論レベルの形式であると述べ、無標形式との呼応関係を認めている。

本稿では、正解データ作成の作業レベルでは呼応を単語－形態論レベルの形式で捉える。しかし、一方で呼応が文の中で他の一定の単語と結合しつつ機能している述語－構文論レベルの形式であることも念頭に置く。

3　正解データ作成

3.1 調査対象語とした呼要素

正解データを作成するにあたり、いわゆる陳述副詞の中から「きっと」「おそらく」「たぶん」「決して」の 4 語を調査対象語として選定した。これらの調査対象語を呼要素と仮定し、調査対象語が出現する文中で応要素とみなせるものにタグ付けを行った。

調査対象語とした「きっと」「おそらく」「たぶん」は、(工藤 1982)による 84 作品の調査結果において

サンプル数が多かった上位 3 語である。「決して」はワープロソフトの校正支援機能によって否定との呼応関係を指摘される認知度の高い陳述副詞である。そこで、「決して」に対する応要素の抽出結果を見ることが、作成した正解データや抽出結果の妥当性を判断する一つの基準になると考え、調査対象語に加えた。

3.2 使用したコーパス

正解データ作成のために、1997 年から 2000 年までの読売新聞、日経新聞、毎日新聞のコーパスを使用した。これらのコーパスを一文毎に区切って形態素解析を行った後、それぞれの調査対象語を含む文を 1200 文ずつ無作為抽出した。

3.3 除外した文

無作為抽出した各 1200 文の中から、以下のものを除外した[1]。

①本のタイトルや見出し内の表現
②倒置　〔例 1〕大変なんだよ、きっと。
③形態素解析の誤りによって、呼要素として抽出されたもの　〔例 2〕どきっとさせられる。
③呼要素と仮定した副詞の一語文
④呼要素と仮定した副詞の後ろが省略された文　〔例 3〕来年にはきっと……。

3.4 作成基準

正解データは、以下の基準で応要素とみなせる箇所にタグを付与することで作成した[2]。

①応要素の「応要素の核部分」「応要素の周辺部分」に分けてタグを付与する

〔例 4〕あたたかさはたぶん、ＣＭの製作者が小錦の人柄と個性と才能に敬意をはらっているから
<核>だろう</核> <周辺>と思う</周辺>。

例 4 では、「だろう」に「応要素の核部分」、「と思う」に「応要素の周辺部分」のタグを付与する。「だろう」「でしょう」などに続く不確実表示用法の文末思考動詞(森山 1992)の「と思う」やそれに準ずる「と考える」などに「応要素の周辺部分」のタグを付与する。

②いわゆる複語尾や助詞や活用形がつかない無標形式も応要素と認め、タグを付与する

〔例 5a〕来夏は、おそらく日本バレーボール協会の強化費で遠征できるだろう。
〔例 5b〕来夏は、おそらく日本バレーボール協会の強化費で遠征できるφ。

例 5a、例 5b はいずれも推量を表す文である。例 5a の文末が「だろう」のある有標形式なのに対し、例 5b の文末は無標形式である。本稿では、文末辞がつかないことも一つの形式であるとする(工藤 1982)の立場を承け、応要素が無標形式になっている文にはゼロ要素にタグを付与する。

③「応要素の核部分」のタグは機能語に付与する

ただし、活用形によっては機能語と内容語に分けると過分割になり、意味をなさなくなるものがある。

〔例 6〕たぶん、これも老若の実力派をまんべんなく集めようとした作家編成に起因していよう。

このような場合、上記例 6 のように「う」のみではなく「いよう」全体にタグを付与する。

④応要素の「応要素の核部分」のタグは、複数のモダリティ階層(益岡 1991)に渡らない範囲に付与する

ただし、文末辞が複数のモダリティ階層に渡る場合は、最小の範囲で複数の階層にタグ付けする。

〔例 7〕こうしたまねっこ行動からは決して失敗を告白し、それを成功の糧とする考え方は芽生えなかった。

例 7 の「なかっ」は「みとめ方のモダリティ」に、「た」は「テンスのモダリティ」に相当する。この場合、連続した「なかった」に対して「応要素の核分」のタグを付与する。

⑤一つの呼要素の係り先が複数ある場合、該当箇所すべてにタグを付与する

〔例 8〕おそらく多くの途上国では、検査体制が不十分で、自分のＨＩＶ感染を知らない人が多いだろうし、たとえ感染が分かっても治療を受けられないだろう。

4　作成した正解データの分析

4.1 タグ付与結果の傾向　－応要素の核部分－

各々の「応要素の核部分」としてタグ付与された表現のうち、出現頻度 5 以上の表現を表 1 に示す。表 1 の［出現頻度］は無作為抽出した 1200 文中での当該応要素の出現数である。また、［割合］は［出現頻度］を 4 年分の新聞記事コーパス全文(18,865,953 文)における当該応要素の全出現数で割った値である。出現頻度が高い表現は、どのような文においても高頻度である可能性がある。そのような場合、分析の際に補正が必要になる。そこで、4 年分の新聞記事コーパス全文中の全出現数に対する割合を求めた。

以下、「応要素の核部分」としてタグ付与された表現の傾向について述べる。「決して」に対する「応要素の核部分」としてタグ付けされた表現では、「ない」「なかった」が圧倒的多数を占め、否定表現が応要素となるという予想通りの結果となった。否定表現ではないものには「難い」などがある。これらは否

[1] 「決して」を含むものから 12 文、「たぶん」を含むものから 88 文、「おそらく」を含むものから 6 文、「きっと」を含むものから 78 文を除外した。
[2] 以下、例文中の＿＿は呼要素と仮定している調査対象語、＿＿は応要素としてタグ付与対象となる箇所を示す。

表 1　応要素の核部分

たぶん	出現頻度	割合(%)	おそらく	出現頻度	割合(%)	きっと	出現頻度	割合(%)	決して	出現頻度	割合(%)
‹ゼロ要素›	542	—	‹ゼロ要素›	465	—	‹ゼロ要素›	528	—	ない	876	0.05
だろう	310	0.18	だろう	340	0.21	だろう	227	0.13	なかった	129	0.04
でしょう	79	0.22	でしょう	74	0.15	でしょう	106	0.21	なく	65	0.02
ではないか	24	0.04	であろう	69	0.51	はず	78	0.23	ず	46	0.02
であろう	23	0.17	ではないか	60	0.10	に違いない	58	0.81	ません	43	0.05
はずだ	13	0.08	に違いない	37	0.51	はずだ	26	0.15	まい	9	0.11
にちがいない	10	0.97	はずだ	14	0.08	はずです	15	0.53	ぬ	6	0.02
じゃないか	9	0.11	たろう	11	1.07	にちがいない	10	0.97			
かもしれない	8	0.02	にちがいない	11	0.52	であろう	7	0.05			
に違いない	8	0.11	かもしれない	9	0.02	では	7	0.00			
はず	7	0.02	ではなかろうか	9	0.61	かもしれない	6	4.26			
かもしれません	5	0.09	はず	9	0.03	に違いありません	6	0.01			
たろう	5	0.16	ではないだろうか	8	0.15						
ではないだろうか	5	0.09	はずである	8	0.42						
らしい	5	0.01	じゃないか	6	0.07						
			まい	6	0.08						
			では	5	0.00						
			ではあるまいか	5	1.02						
			らしい	5	0.01						

定の「『相当』『準用』形式」(工藤 1982)で、その部分だけを見ると否定ではないが、文全体を見ると否定の意味となる表現である。

〔例 9〕演出や歌唱など、それぞれを単独で評価すると決して満点とは言い難いのだが、互いを巧みに結びつけ、ひとつの有機体に仕上げる手際よさはまさに伝統のなせる業である。

次に「たぶん」「おそらく」「きっと」に対する「応要素の核部分」としてタグ付けされた表現の傾向について述べる。タグ付けされた表現のうち最も多かったのは‹ゼロ要素›で、それぞれ付与したタグの約半数を占めた。[出現頻度]のみで見ると、‹ゼロ要素›以外で多いのは「だろう」「でしょう」「ではないか」「であろう」「はず」「はずだ」「に違いない」「かもしれない」で、これは「たぶん」「おそらく」「きっと」に共通する傾向である。だが、[出現頻度]と[割合]の双方で見ると、「たぶん」に対する応要素として「にちがいない」、「おそらく」に対して「たろう」「ではなかろうか」「にちがいない」、「きっと」に対して「かもしれない」「にちがいない」「に違いない」「はずです」が上位に挙がる。しかし、「応要素の核部分」を見る限り、「たぶん」「おそらく」「きっと」の 3 語に大きな違いは現れない。

4.2 タグ付与結果の傾向　ー応要素の周辺部分ー

「たぶん」「おそらく」「きっと」については、「応要素の周辺部分」としてタグ付与した表現がある。出現頻度 3 以上のものを表 2 に示す。

「と思う」「と思います」「ね」「な」など出現頻度の高い表現が 3 語に共通して見られるのは「応要素の核部分」と同様である。一方、調査対象語によって傾向が分かれる点もある。「たぶん」に対する「応要素の周辺部分」は「と思う」が頻出するが、「おそらく」の方では「との見通しを示した」「と考えられる」「と推測する」、出現頻度が 3 未満のため表 2 には掲載のない「とみられる」「との見解を示す」「と想像する」「と予想する」「と判断する」など、自らの考えを示す表現が見られる。また、「きっと」に対する「応要素の周辺部分」では、「と信じている」「と

表 2　応要素の周辺部分

たぶん	出現頻度	おそらく	出現頻度	きっと	出現頻度
と思う	77	と思う	42	と思う	67
と思います	35	と思います	26	よ	49
ね	23	と思われる	16	と思います	43
よ	14	な	7	ね	33
な	11	と思われます	6	な	14
と思いますよ	6	ね	6	と思いました	5
と思うんです	4	との見通しを示した	4	と思った	5
と思った	4	と考えられる	3	と信じている	5
とは思う	3	と思った	3	なあ	5
と思われます	3	と推測する	3	わ	5
と思われる	3			ぞ	4
				と思っている	4
				と信じています	4
				ことと思います	3
				と確信している	3
				と思うんです	3
				と思っています	3
				よね	3

確信している」、表 2 には掲載していない「と期待する」「のになぁ」など、良いことを信じたり期待したりする表現が見られる。

5　抽出結果との比較検討

(山本ほか 2004)では、語や文字列の類似度を測る尺度を用いて大規模コーパス[3]から呼応表現の抽出実験を行っている。この実験では 7 種類の尺度を使って呼応表現を抽出し、各尺度によって得られた結果の上位 500 件をプーリングすることでデータを作成している。

抽出結果をプーリングしたデータには、「きっと」に対する応要素として「思うの」「あるの」などが含まれている。ここで「ある」「思う」は応要素としては不要部分とも判断できるが、手作業で作成した正解データには入らなかった「の」を発見することができる。また、「決して」のデータに含まれている「ございません」も手作業での正解データでは得られなかったものである。

6　まとめ

本稿では人手作業によって、「たぶん」「おそらく」「きっと」「決して」の 4 種の調査対象語に対して、各 1200 文の正解データを作成した。その過程において、以下の知見が得られた。

(1) データ作成における人手作業と自動抽出

人手作業は自動抽出に比べてかかる労力が高く作業量は限られるが、正確なデータを作成することができる。人手作業で一つ一つのデータを観察していくことにより、「おそらく」と「と見通しを示す」「とみられる」「と考えられる」「と推測する」「と想像する」などの自らの考えを示す表現が共起しやすく、「きっと」と「と信じる」「と確信する」「と期待する」など良いことを信じたり期待したりする表現が共起しやすいなどの現象の発見が可能であった。これらの表現は自動抽出により作成したデータにも含まれているが、自動抽出の結果のみからこれらの現象を考察することは難しい。

一方、コーパスからの自動抽出によってデータを作成する方法(山本ほか 2004)では、比較的労力をかけずに網羅的に大容量データを扱うことが可能になる。自動抽出の場合、不要な表現を大量に取ってしまうことで精度に問題が出る点は否めないが、「～の」「ございません」などの新たなバリエーションの呼応表現を発見的に抽出できる。

このように人手作業にも自動抽出にもそれぞれ長所と短所がある。大規模コーパスからの自動抽出によってデータの表現バリエーションを広げ、網羅的に呼応表現を抽出することと、人手作業によって正確なデータを作成することの二つは両輪である。質の高い網羅的なデータの作成を目指すには、人手作業と自動抽出の双方の精度を上げていくことが重要である。

(2) 自動抽出の今後

「たぶん」「おそらく」「きっと」に対する「応要素の核部分」の約半数を占めた〈ゼロ要素〉や、「決して」に対する「難かった」などの「『相当』『準用』形式」は、呼応を述語－構文論レベルの形式と見なければ記述できないものである。抽出の自動化を考える場合、これらを類似尺度のみによって抽出することは難しい。構文規則などを用いた新たな抽出規則を作成する必要がある。これらは今後の課題としていきたい。

謝辞　本稿を纏めるにあたり、毎日新聞社、読売新聞社、日経新聞社の新聞記事データを使用させて頂きました。感謝致します。

[3] 全体で 38,875,937 文の 38 年分の新聞記事コーパス。内訳は、読売新聞 15 年分、日経新聞 11 年分、毎日新聞 12 年分である。

参考文献

大野晋 ： 係り結びの研究, 岩波書店 (1993).
橋本進吉 ： 国文法体系論, 岩波書店 (1959).
木田敦子,乾裕子,神崎享子,高梨克也,井佐原均 ： 構文論から見た対話 －円滑な話者交替を可能にする構文構造－, 第 33 回人工知能学会 言語・音声理解と対話処理研究会資料, SIG-SLUD-A102, pp.33-38 (2001).
Atsuko Kida, Eiko Yamamoto, Kyoko Kanzaki, and Hitoshi Isahara : Extraction and Verification of KO-OU Expressions from Large Corpora. *ACL-03 Companion Volume to the Proceedings of the conference*, pp.169-172 (2003).
工藤浩 ： 叙法副詞の意味と機能 －その記述方法をもとめて－,国立国語研究所報告 71 研究報告集 3 (1982).
益岡隆志:モダリティの文法,くろしお出版 (1991).
松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 ： 形態素解析システム『茶筌』version 2.2.9 使用説明書, 奈良先端科学技術大学院大学 松本研究室 (2002).
森山卓郎 ： 文末思考動詞「思う」をめぐって －文の意味としての主観性・客観性―, 日本語学 Vol.11 No.8, pp.105-116 (1992).
山田孝雄 ： 日本文法論, 宝文館 (1908).
山本英子, 木田敦子, 神崎享子, 井佐原均 ： コーパスからの呼応表現自動抽出手法の評価,言語処理学会第 10 回年次大会発表論文集,D1-2 (2004).