一般論文

大規模コーパスからの呼応表現抽出

Extraction of KO-OU Expressions from Large Corpora

木田　敦子*　山本　英子　神崎　享子**　井佐原　均****

By Atsuko KIDA, Eiko YAMAMOTO, Kyoko KANZAKI and Hitoshi ISAHARA

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、述語が目的語などに後置されるため、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。述語が後置される文法構造を持つ日本語において、この漸進的文理解を可能にする要因の一つに、「呼応関係」という構文構造があると考えられる。

「呼応関係」とは、「決して行かない」の「決して」と「ない」のように、「先行する一定の語に応じて後ろに特定の形が来る」(『岩波国語辞典』より）関係である。これに対して、「共起関係」は、「赤い花」の「赤い」と「花」のように、二つの語が同一文内に出現する関係である。共起関係には出現順序に制約はないが、呼応関係には制約がある。本稿では、以下、「先行する一定の語」を「呼」要素、「後ろにくる特定の形」を「応」要素と呼ぶ。

中世以前の日本語には、係助詞と文末の活用形とが形態的な呼応関係を持つ係り結びの用法があった。古語では、係り結びの機能によって、係助詞が後続要素を予告する働きを担っていたと見ることもできる。大野[1]は、係り結びが消滅した現代語においても、ある種の副詞が古語の係助詞と似た役割を果たしており、文の行く手を予告する働きを持っていることを指摘している。実際、現代語にも「しか～ない」「決して～ない」などの呼応関係が存在する。我々は、このような表現を収集して、呼応表現データを作成することを試みている。従来、このような客観的な基準を用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。本研究で作成する呼応表現データは、対話処理システムに求められる漸進的な文理解[2]や文予測のための基礎データとして役立つと考えられる[3]。また、このデータは、構文解析の曖昧性解消や係り受け関係決定の補助情報としても役立つと考えられる。

そこで我々は、コンピュータによる処理を前提とした電子テキストの集合体であるコーパス[7]から呼応表現を自動抽出し、データを作成することを目指している[4),5),6)]。本稿では、コーパスから呼応表現を抽出する手続きと、抽出した呼応表現データの分析結果について報告する。

2．本研究の位置付け

大野[1]は、古語の係助詞に代わって「ある種の副詞」が、「時間的に線状的に発展し連続していく言語表現の早い部分で、一文の行く手、肯定か否定か疑問かなどを予告しておこうとする」役割を果たしていることを示唆している。

また、益岡[9]のモダリティ論では、文を階層構造と呼応関係を持つものと捉え、以下のような呼応関係を挙げている（表−1)。

表−1　益岡による呼応関係

「呼」要素	「応」要素
ねえ、おい	ね、よ
ぜひ、なんて	て下さい、なあ
たぶん、どうも、いったい	だろう、らしい、か
むかし、かつて、もうすぐ	た
決して、必ずしも	ない

＊言語情報研究室　＊＊独立行政法人通信総合研究所

大野[1]の現代語における係り結びに代わる現象の指摘は大変貴重なものだが、具体的な例を挙げての説明はない。また、益岡[9]は、呼応関係にある「呼」要素のグループ「たぶん・どうも・いったい」に対して「応」要素のグループ「だろう・らしい・か」を挙げているが、個々の語がそれぞれどの語と呼応関係にあるのかは記述していない。また、挙げられている要素の数が少なく、挙げられている「呼」要素と「応」要素の客観性が保証されていない、などの弱点がある。そこで我々は、大規模な電子化コーパスから自動的に呼応関係を抽出することで、客観的かつ実用に耐える規模の呼応表現データを作成することを目指す。

3．方法と調査対象

(1) 方法

補完類似度は文字認識（パターン認識）の分野で用いられている類似尺度である。出現パターン（ベクトル）の包含状況を測ることによって関係を推定するため、パターンの包含関係の抽出に強い。

山本・梅村[13]は、出現パターンの包含関係に強い補完類似度を用い、コーパスから事象の一対多の関係を抽出する実験を行っている。

呼応関係の出現パターンを見ると、「応」要素は「呼」要素より頻繁に出現する。補完類似度は出現パターンの包含関係を測るものなので、「応」要素の出現パターンが「呼」要素の出現パターンを包含するかどうかを測り、その二つが包含関係にあれば、「呼」要素の出現パターンは「応」要素の出現パターンと重なる部分が多く、補完類似度は高い値を保持する。

呼応関係は「呼」要素がいくつかの「応」要素を持つ一対多関係である。一方、出現パターンに関しては、「応」要素が「呼」要素を包含する一対多関係になる。そこで、我々は補完類似度を用い、出現パターンにおいて「応」要素が「呼」要素を包含する関係にあることを利用して、呼応表現の抽出実験を行った。本研究では、調査対象語を呼応関係の「呼」要素と仮定し、補完類似度を用いて調査対象語に対する「応」要素を抽出することを試みる。

(2) 対象データ

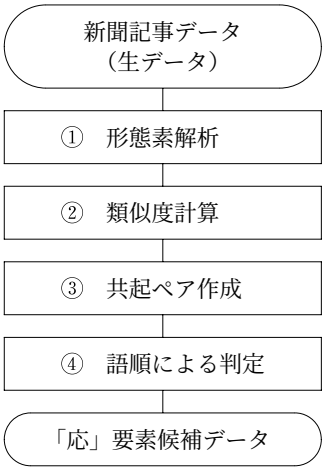
本稿では、いわゆる副助詞、係助詞と副詞の一部である以下の語を調査対象語とした。これらを呼応関係の「呼」要素と仮定して調査を進めた。

［調査対象語］
は・も・こそ・さえ・しか・決して・ぜひ
まるで・もし・きっと

調査データには、1991 年から 2000 年までの毎日新聞、日本経済新聞、読売新聞の新聞記事データを使用した。データの規模は、毎日新聞記事データ 10 年分の 10,273,385 文、282,859,867 形態素、読売新聞記事データ 10 年分の 14,938,734 文、466,642,043 形態素、日本経済新聞記事データ 10 年分の 15,565,344 文、461,733,326 形態素、合計 40,777,463 文、1,211,235,236 形態素である。

4．補完類似度を用いた呼応表現の抽出

処理の流れを図−1 に示す。



図−1　処理の流れ

① 新聞記事データ（生コーパス）を一文ごとに区切った後、形態素解析システム茶筌[11]を用いて形態素解析を行う。活用語は原形に変換する。

② 補完類似度によって、固有名詞、普通名詞、数詞以外の全形態素間の類似度計算を行う。類似度計算は、新聞の種類別に 10 年分の記事データに対して行う。

③ 類似度計算の結果から、調査対象語を含むペアを抽出する。これを共起ペアと呼ぶ。

④ 3で得られた共起ペアに対して、語順による判定を行う。共起ペアが、「呼」要素－「応」要素の順で出現している文の数が、「応」要素－「呼」要素の順で出現している文の数より多ければ、呼応候補ペアであると判定する。ここで得られた呼応候補ペアのうち、調査対象語ではない方の要素が「応」要素候補となる。

表－2は、各「呼」要素に対して抽出された「応」要素候補の出現数を示したものである。出現数が最も多い「は」では18,590個の要素が、最も少ない「ぜひ」では1,977個の要素が抽出された。抽出された「応」要素候補には、「決して（副詞-一般）－平たん（名詞-形容動詞語幹）」「決して（副詞-一般）－偶然（名詞-形容動詞語幹）」のような呼応表現と関係ないものも多く含まれている。類似度が低くなるにつれ、このような呼応表現ではないと判断できるものが増えていく傾向が見られる。そこで、それぞれの「呼」要素に対する「応」要素候補として抽出されたものの中から、類似度順に上位10位までを選び分析を行った。

5．得られた「応」要素の分析

本章では、「応」要素候補の分析結果の中から、「呼」要素「きっと」に対する「応」要素候補の分析結果を取り上げて詳細を述べる。表－3の“「応」要素候補”欄は、「きっと」に対する「応」要素として抽出された候補から、類似度順に上位10位までを選んだものである。それぞれの「応」要素候補に対して、目視で実例を100例ずつ観察し、

・単独で「応」要素になるもの　○
・組み合わせで「応」要素になるもの　△
・「応」要素ではない可能性があるもの　×

の判定を行った。判定の結果は、表－3の“分類”欄にそれぞれ上記の○×△の記号で付与した。迷うものは判定を保留し、空欄とした。

判定の結果、「呼」要素「きっと」に対する「応」要素候補として得られたものの中には、“「応」要素ではない可能性があるもの（×）”と判定できるものは見られなかった。そこで、以下では“単独で「応」要素になるもの（○）”、“組み合わせで「応」要素になるもの（△）”について詳しく述べる。

表－2　各「呼」要素に対して抽出された「応」要素候補の出現数

	は	も	さえ	しか	こそ	もし	きっと	まるで	決して	ぜひ
候補数	18,590	9,597	3,712	3,530	6,094	8,222	2,865	4,673	4,039	1,977

表－3　抽出実験から得られた「きっと」に対する「応」要素候補上位10語

「呼」要素	「応」要素候補	類似度	「応」要素候補・出現数	分　類
きっと	う（助動詞／不変化型）	0.004726	1,333,016	△
きっと	だ（助動詞／特殊・ダ）	0.004180	15,642,869	△
きっと	と（助詞－格助詞－引用）	0.003722	10,193,119	
きっと	て（助詞－接続助詞）	0.003030	19,840,812	
きっと	思う（動詞－自立／五段・ワ行促音便）	0.002636	647,692	
きっと	です（助動詞／特殊・デス）	0.002215	1,042,087	△
きっと	ない（助動詞／特殊・ナイ）	0.002002	5,492,225	△
きっと	ます（助動詞／特殊・マス）	0.001957	1,417,320	
きっと	の（名詞－非自立－一般）	0.001953	3,855,773	
きっと	はず（名詞－非自立－一般）	0.001560	131,976	○

(1) 単独で「応」要素になるもの

表−3に挙げた「呼」要素「きっと」に対する「応」要素候補のうち、「はず」を“単独で「応」要素になるもの”と判定した。

（例１）　広大な海辺で見る作品はきっと魅力的に映るはずだ。

だが、これ以外には、単独で「きっと」の「応」要素となると判定できるものが見られなかった。“単独で「応」要素になるもの”と判定できる要素が少ないことは、「きっと」に対する「応」要素候補以外のデータにも共通して見られる傾向である。

なお、「思う」は“単独で「応」要素になるもの”と判定しなかった。「きっと」は「だろう」などの推量の意味を持つ要素と呼応する傾向にある。そのため、主観的な意味の強い「思う」を「応」要素と判定したくなる。だが、実例をあたると、多くの場合、「思う」は引用の助詞「と」に後続する形で出現していることが観察できる。

（例２）　厳しい道だが、後に続く選手はきっとたくさん出てくると思う。

例２は、「きっとたくさん出てくるだろうと思う」「きっとたくさん出てくるはずだと思う」に言い換え可能である。そして、このように言い換えた場合、「きっと」と「だろう」、「きっと」と「はず」に呼応関係が認められると判断されることになる。このように考えると、例２の場合も、呼応関係にあるのは「きっと」と「思う」ではなく、「きっとたくさん出てくるϕと思う」の「きっと」と「ϕ」であると見ることもできる。本稿では「ϕ」を「応」要素と見るか否かについてはこれ以上議論せず、重要な問題であるとの指摘のみにとどめたい。

(2) 組み合わせで「応」要素になるもの

「だ」「う」「です」「ない」を“組み合わせで「応」要素になるもの”と判定した。「だ」は「だろ」の原形である。これと「う」が組み合わさり、「だろう」になる。

（例３）　宇宙飛行士たちは、きっとこのような漆黒の空間を見たのだろう。

同様に、「です」「う」が組み合わさって「でしょう」に、「違い」「ない」が組み合わさって「違いない」になる。このように前後の語とのまとまりで見ることによって、呼応関係が認められる「応」要素候補がある。

（例４）　落語家にならなかったらきっと、長屋の住民の目線で生きる庶民的な弁護士になっていたでしょう。

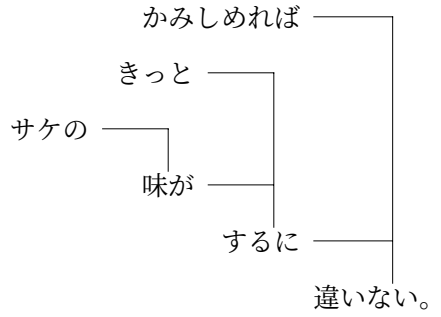
（例５）　酸いも甘いもかみ分ける武双山ならきっと正しい方向性を見つけるに違いない。

６．呼応表現データの実用可能性

(1) 曖昧性解消・係り受け関係決定

「呼」要素と「応」要素の関係を記述した呼応表現データは、構文解析の曖昧性解消や係り受け関係の決定に役立つ。

「かみしめれば、きっとサケの味がするに違いない。」を構文解析システム KNP[8]を用いて解析すると、下記のような結果が得られる。

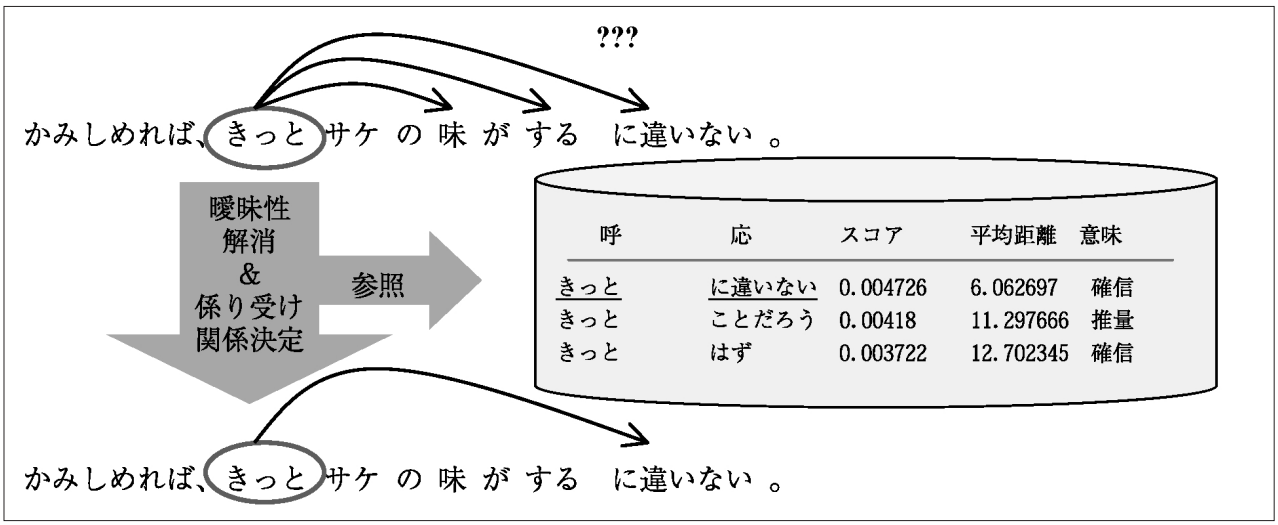


ここで「きっと」は「する」に係ると誤って解析されている。このような場合、呼応表現データは正しい係り受け関係を決定するための補助情報となり得る（図−2)。

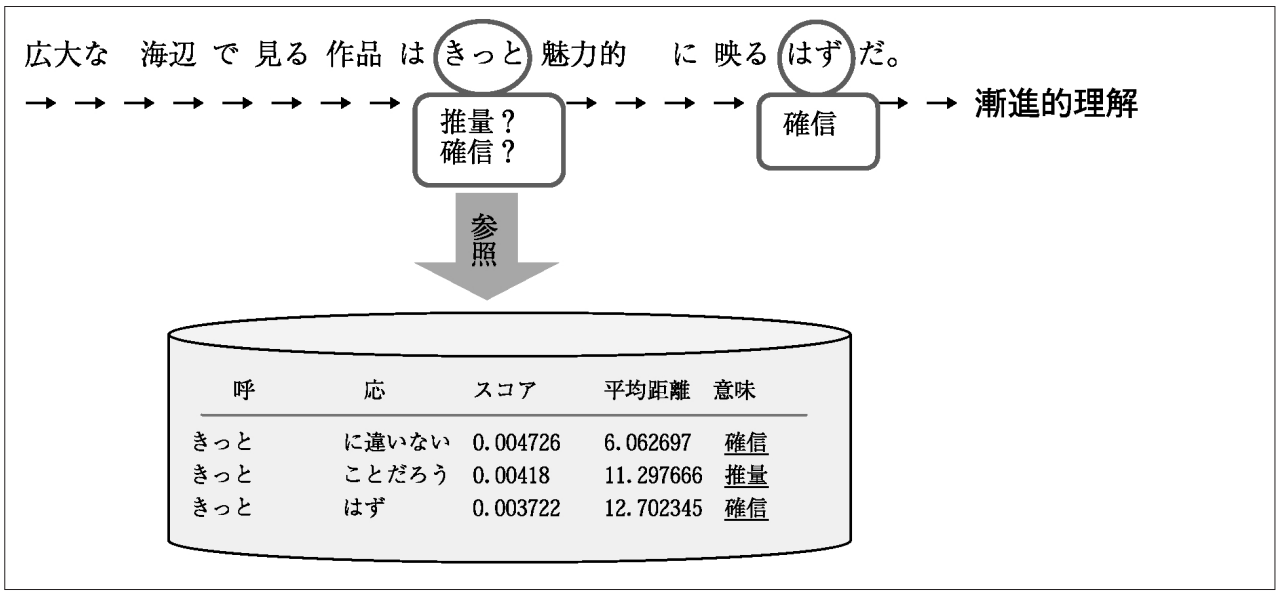
(2) 漸進的文理解

呼応表現データをシステムに組み込めば、人が進行していく文を漸進的に理解していくように処理を進めることが可能になる。たとえば「呼」要素「きっと」が現れた段階で、システムは後続要素に推測や確信の意味を持つ語が現れることが予測できる（図−3)。これは、話し言葉に見られる言い差しの文や、Web 上の文章などに見られる完全ではない文の処理に対応できる柔軟な技術の開発につながると考えられる。

また、呼応表現によって文末表現の予測が可能になることから、複数人の会話において話者が交替した箇所の推測ができると考えられる。これは会議議事録などの自動書き起こしシステムのための基礎技術として利用できる可能性がある。



図−2 曖昧性解消



図−3 前進的文理解

7．意味情報付与の必要性

呼応表現データを文の漸進的理解に用いるなら、データに意味情報が付与されていることが必要である。データに意味情報が付与されていれば、「呼」要素が現れた段階で、呼応表現データを参照することによって文内容を十分につかんだり推測したりすることが可能になる（図−3)。だが、第４章の処理で得られた「応」要素候補に意味情報を付与するのは難しい。形態素解析によって一形態素ごとに分割されており、各要素が短すぎるためである。

通常、述語は文意決定に重要な役割を果たしている。日本語の文の述語は、動詞に続く助詞、助動詞の連続によって成り立っている場合が多い。この連続部分に含まれる助詞、助動詞は機能語である。それ故に、助詞・助動詞の中には単独で見ても、意味が決定できないものがある。また、他の語と組み合わさることによって意味が変わるものもある。

表−3に挙げた「応」要素候補「だ」は「今日は雨だ」という文では断定の意味になる。一方、「明日は晴れるだろう」という文では推量の意味になる。また、「ない」は否定を表す助動詞である。だが、「かもしれない」「に違いない」という語とのまとまりでは、否定の意味は消える。そして、語のまとまり全

体で、推量、確信の意味になる。

呼応の「応」要素である判断できるものには、述語が多く見られる。したがって、前述した述語の語構成の性質上、意味情報付与の前段階で複数要素を組み合わせる作業が必要であると考えられる。

８．おわりに

以上、大規模コーパスからの呼応関係の抽出方法とその分析結果について報告した。

本稿では、補完類似度を用いた大規模コーパスからの呼応表現データ作成の手順、呼応表現データの分析結果、呼応表現データの有用性について述べた。

また、意味情報付与の必要性について述べ、意味情報を付与する前に、複数要素を組み合わせる作業が必要であることについて論じた。現在、我々は複数要素の組み合わせを自動生成する課題に取り組んでいる。現段階の処理では「応」要素として得られるのは「はず」「だ」のように一形態素のみだが、複数要素の組み合わせ自動生成が実現できれば「に違いない」「でしょう」「だろう」「かもしれない」などの自動抽出が可能になる。複数要素の組み合わせの自動生成は、今後の課題としていきたい。

参考文献

１）大野晋：係り結びの研究，岩波書店，1993.

２）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－A 102, pp. 33－38，2001.

３）木田敦子，乾裕子，高梨克也，井佐原均：文構造の漸進的予測を可能にする日本語の諸特徴の分析，言語処理学会第８回年次大会発表論文集，pp. 65－68，2002.

４）木田敦子，山本英子，井佐原均：後続要素を予告する表現の分析，情報処理学会研究報告，NL－152－20，pp. 137－143，2002.

５）木田敦子，山本英子，神崎享子，井佐原均：呼応関係を産み出す構文手がかり，言語処理学会第９回年次大会発表論文集，pp. 23－26，2003.

６）Kida, A., Yamamoto, E., Kanzaki, K. and Isahara, H.: Extraction and verification of KO-OU expressions from large corpora, ACL-03 Companion Volume to the Proceedings of the conference, pp. 169-172, 2003.

７）後藤斉：言語理論と言語資料－コーパスとコーパス以外のデーター，日本語学４月臨時増刊号，Vol. 22，pp. 6－15，2003.

８）黒橋禎夫：日本語構文解析システムKNP version 2.0 b 6，京都大学大学院情報学研究科，1998.

９）益岡隆志：モダリティの文法，くろしお出版，1991.

10）益岡隆志、田窪行則：基礎日本語文法―改定版―，くろしお出版，1992.

11）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，2002.

12）山田孝雄：日本文法学概論，宝文館，1936.

13）山本英子，梅村恭司：コーパス中の一対多関係を推定する問題における類似尺度，自然言語処理，Vol. 9，No. 2，pp. 45－75，2002.

謝辞：本稿を纏めるにあたり、検索ツールteaの使用を許可して下さいました通信総合研究所の内山将夫氏に感謝致します。また、毎日新聞社、読売新聞社、日本経済新聞社の新聞記事の電子化データを使用させて頂きました。感謝申し上げます。