ひらがな列を手がかりとした文末機能表現の自動抽出

松木 久幸　　佐藤 理史
名古屋大学工学部　電気電子・情報工学科　　名古屋大学大学院工学研究科　電子情報システム専攻
{h_matuki, sato}@nuee.nagoya-u.ac.jp

1　はじめに

日本語には、文末の用言に接続して複雑な述語を作る表現—本稿では文末機能表現と呼ぶ— が存在する。文末機能表現は、ヴォイス、テンス、アスペクト、モダリティなどの多様な情報を、用言や文に付加する働きを持つ。たとえば、例文

(1) 原稿を書きはじめているかもしれません。

において、「はじめる」は「動きの開始 (アスペクト)」を、「ている」は「動きの継続 (アスペクト)」を、「かもしれません」は「可能性 (モダリティ)」を表す。これらの要素を把握することは、文が伝える情報を正確に把握するために欠かすことができない。

我々は、このような文末機能表現の体系的整理とその工学的利用を目指し、文末機能表現辞書の編纂を、次のような手順で進めている。

1. 文末機能表現を整理するための大枠 (構造モデル) を定める。
2. これと並行して、辞書の見出し語リストを暫定的に定める。
3. コーパスに現れる実例を調査し、構造モデルおよび見出し語リストを修正する。

本稿では、この手順のステップ3の実施に必要な実例を、コーパスからどのように収集するかについて述べる。そもそも辞書を作る目的は、あるクラスに属する表現—我々の場合は、文末機能表現—の範囲を厳密に定めるためである。つまり、辞書の編纂過程では、対象とする表現クラスは厳密には定まっておらず、そのようなクラスの実例を機械的に集めることは、明らかに不良設定問題となる。

本稿の構成は以下の通りである。まず、2節で、我々が採用した構造モデルについて述べる。3節では、コーパスから文末機能表現の候補をどのように抽出するかについて述べ、4節で、抽出結果を示す。

2　文末述語表現の構造モデル

本稿では、用言とそれに後続する文末機能表現を併せて、文末述語表現と呼ぶ。我々は、まず、次のような手順で、文末述語表現のおおまかな構造を定めた。

1. 文 (実在する文または作例) において、直感に従って、文末の述語表現の範囲を定める。
2. 切り出された文末述語表現内に境界を認定し、それらをいくつかの部分に分割する。
3. こうして得られた部分を整理・グループ化することにより、文末述語表現のモデルを定める。

なお、準拠する文法体系 (品詞体系、活用体系等) として、益岡・田窪文法 [2] を採用した[1]。

たとえば、

(2) 噂が話されはじめているかもしれませんね。

という文において、下線部を文末の述語表現と認定することに異論はないだろう。この文末述語表現において、

(3) 話さ–れ | はじめて | いる ‖ かもしれません ‖ ね。

のように、3種類の境界を考える。

1. 境界 ‖ は文末となりうる境界である。その直前は、基本形、タ形となる。
2. 境界 | は文節末となりうる境界である。その直前は、連用形、テ形となる。
3. 最後の境界–は、そこで切るとその左が文節として自立しなくなる境界である[2]。

このような3種類の境界の認定に基づき、文末述語表現の構造を図1に示すような形でモデル化した。こ

[1] 益岡・田窪文法は、助動詞として認めるものが少なく、活用体系によって抽象化される範囲が大きいので、考慮すべき構成要素の数が少なくなり、整理において見通しが立てやすい。

[2] 益岡・田窪文法では、未然形という活用形は設定されず、「〜(ら) れる」は接辞扱いである。

$$
\begin{aligned}
\text{文末述語表現} &= \text{用言 U} + \left\{ \begin{array}{l} \mid \text{準用言 U} \\ \parallel \text{助動詞 U} \end{array} \right\}^{*} + \{\parallel \text{終助詞 U}\}^{*} \\
\text{用言 U} &= \text{用言} + \{- \text{述語系接尾辞}\}^{?} \\
\text{準用言 U} &= \text{準用言} + \{- \text{述語系接尾辞}\}^{?} \\
\text{助動詞 U} &= \text{助動詞} + \{- \text{述語系接尾辞}\}^{?} \\
\text{終助詞 U} &= \text{終助詞} + \text{終助詞}^{*} \\
\text{用言} &= \left\{ \begin{array}{l} \text{動詞} \\ \text{形容詞} \\ \text{名詞} + \text{判定詞} \end{array} \right\}
\end{aligned}
$$

注：動詞は「名詞+する」を含む

図 1: 文末述語表現の構造モデル

のモデルにおいて、肩付きの記号 '*' は任意個の並び、'?' はあってもなくてもよいことを表す。境界との対応を明確にするために、境界を含んだ形で示してある。なお、U はユニットの略である。

このモデルに従えば、

1. 境界 ‖ の右に現れるものは、助動詞[3]か終助詞のいずれかである。
2. 境界 | の右に現れるものは、準用言[4]である。
3. 境界–の右に現れるものは、述語系接尾辞である。

すなわち、3 種類の境界を定めれば、それによって切り出される部分がどのタイプの構成要素となるのか、ほぼ定まる。

現在編纂中の辞書では、用言 U、準用言 U、助動詞 U、終助詞 U に分類される表現を見出し語として採用する[5]。但し、用言 U は、用言の部分を品詞レベルに抽象化したものを見出し語として採用する。すなわち、見出し語を定めることは、次のリストを作成することと等価である。

1. 用言 U の網羅的リスト (その中核は、述語系接尾辞の網羅的リスト)
2. 準用言 U の網羅的リスト
3. 助動詞 U の網羅的リスト
4. 終助詞 U の網羅的リスト

我々は、主に、桝田らのシソーラス [1] を再整理して、暫定的な見出し語を定めた。現時点での見出し語数を表 1 に示す。

[3]益岡・田窪文法の助動詞に対応する。

[4]我々が独自に導入したクラスである。益岡・田窪文法の連用形補助動詞、テ形補助動詞、および、連用形・テ形接続の形容詞 (「やすい」、「ほしい」等) が含まれる。

[5]末尾の活用形は抽象化し、基本形のみを収録する。

表 1: 見出し語数一覧

タイプ	述語系接尾辞		計
	なし	つき	
用言 U	3	19	22
準動詞 U	87	576	663
助動詞 U	27	43	70
終助詞 U	24	–	24
合計	141	638	779

3 文末機能表現候補の抽出法

辞書編纂の次のステップは、暫定的に定めた見出し語集合とコーパスとを照合し、重要な表現が未収録となっていないかを確認する作業となる。自然言語の性質上、数学的意味での完全な列挙を追求することは不可能に近い。そのため、「実際に使用されている表現のほとんど (あるいは、大部分) を収録すること」が辞書編纂の現実的な目標となる。

3.1 ひらがな列の利用

ある文が与えられたとき、その文の文末述語表現を定めることは、その左境界を定めることにほかならない (ただし、体言止めの文のように、文末述語表現が存在しない文もある)。採用した構造モデルに基づけば、文末述語表現の先頭は用言である。それゆえ、文を形態素解析し、文中の最後の用言の左境界を文末述語表現の左境界として採用すればよいように思われるが、次のような問題が生じる。

a. 複合辞の認定
   形態素解析器によっては、「食べるかもしれない」を「食べる/かも/しれ (動詞)/ない」、「書くことがある」を「書く/こと/が/ある (動詞)」と解析するものがある。これらの場合、「〜かもしれない」、「〜ことがある」が抽出されないことになる。
b. 文法体系のミスマッチ
   形態素解析器によって、準拠している文法体系が異なる。
c. 形態素解析誤りの存在

そこで、我々は、形態素解析の結果に加えて、文末に現れるひらがな列を手がかりとして利用する方法を採用する。一般に、機能表現は、ひらがなで書かれることが多い。我々はこの事実を拡大解釈し、「文末機能表現はひらがなで書かれる」と仮定する。このような仮定を採用した場合、必ず漢字表記される表現 (た

とえば、「〜可能性がある」) を抽出することはできないが、ひらがなでも書かれることがある表現 (たとえば、「〜ざるを得 (え) ない」) であれば、抽出することが可能である。

具体的には、次の手順に従って、調査対象とする文末形態素列を決定する。

1. 文を形態素解析する。
2. 文末の句点と「？」を除去する。
3. 形態素列を右 (文末) から左に順に調べていき、表記文字列にひらがな以外の文字を含む、最も右の (文末に近い) 形態素を特定する。
4. この形態素の表記文字列が、ひらがな、漢字、カタカナのみから構成されている場合は、その形態素から文末までの形態素列を出力する。そうでない場合は、その文を破棄する。

以下では、この手順で得られる文末形態素列を $M = m_l m_{l-1} \cdots m_1$ のように表記する。添字は末尾からの位置を表す。この抽出手順は、形態素境界を利用しているため、用いる形態素解析器によって、抽出結果は異なることがある。今回は、次の 3 つの形態素解析器を利用する。

a. JUMAN
b. MeCab + UniDic
c. MeCab + IPAdic

3.2 文末機能表現の抽出

こうして得られた文末形態素列 M には、ひらがなで表記された内容語が含まれる場合がある。そこで、この文末形態素列を以下の手順で再度調査し、最終的に抽出する文末述語表現を決定する。

1. 形態素列 M を左から右にスキャンし、それぞれの形態素 m_i に P, F, O のいずれかのラベルを付与する。
2. 形態素列 M を右から左にスキャンし、次の条件を満たす形態素 m_i が見つかったならば、$m_i \cdots m_1$ を文末述語表現として抽出し、手続きを終了する。
   (a) m_i のラベルは P である。
   (b) $m_j (i > j \geq 1)$ のラベルは F である。

ここで、P, F, O の各ラベルは、「用言」、「機能語」、「それ以外」を意味する。

この手順の中核は、ステップ 1 のラベル付与である。これは、次の方針で行なう。

a. 原則として、使用した形態素解析器の文法体系に基づき、付与するラベルを決める。
b. ただし、一部の形態素に関しては、その決定をオーバーライトする。

たとえば、今回使用した 3 つの形態素解析器において、「ある」は動詞であり、これに基づくラベルは P(用言) となる。しかし、「ある」は助動詞相当の複合辞の構成要素となるので、ある条件下ではこれを F に書き換える。書き換えの条件の詳細は、それぞれの形態素解析器に対して異なるが、おおよそ次のようにまとめられる。

1. 複数の形態素から構成される用言 (名詞+だ、名詞+する、形容動詞) を認定し、ラベル P を付与する。(仮想的に、一つの形態素とみなす。)
2. 形式名詞のように働く名詞のラベルを O から F に書き換える。(UniDic に対して適用)
3. 動詞または形容詞で、その直前の形態素が動詞または接尾辞の場合、そのラベルを P から F に書き換える。(UniDic と JUMAN に対して適用)
4. 「ある、ない、いる、なる、できる、いい、よい、する、いう、みる」のいずれかであり、それより左側に P があり、かつ、その間に O がない場合、ラベル P を F に書き換える。

最終的に、抽出した文末述語表現の用言部分を品詞 (最上位レベル) で置き換えたものを作成し、文末機能表現候補として出力する。なお、末尾の活用形は基本形に正規化せず、そのままの形で残す。

4 抽出された文末機能表現候補

先に示した 3 つの形態素解析器を利用して前節の手順を実行し、どのような文末機能表現候補が抽出されるかを調べた。コーパスには、CD-毎日新聞データ集 (2005 年版) の 135 万文を用いた。JUMAN、UniDic、IPAdic を用いた場合、抽出された文末機能表現候補数は、それぞれ、65 万 (14,159)、64 万 (14,169)、62 万 (14,390) である。なお、括弧内の数字は種類を表す。

抽出された表現のうち、文字数の長いもの 42 件を表 2 に示す。ただし、ここでは、いずれかの形態素解析器を用いた場合に、頻度が 3 以上であったもののみに限定した。この表が示すように、用いる形態素解析器によって、得られる結果が異なる。これは、それぞれの形態素解析器が準拠している文法体系が異なることによる。UniDic と IPAdic が準拠している文法体系の差異は比較的小さいが、品詞細分類が異なる語彙も

あり、その影響で、得られる結果が異なってくる。このように、複数の形態素解析器を用いた結果をマージすることで、より多くの文末機能表現候補を、抽出の網にかけることができる。

しかしながら、現在採用している方法には、いくつかの課題がある。その中の最大の課題は、文末の定型表現を十分に拾うことができていないという点にある。現在の方法は、3.1 節で抽出した形態素列 M の中に複数の P(用言) が含まれている場合、単純に、一番右側 (文末に近い) の P を選ぶ方法を採用している。しかしながら、文末のひらがな列に複数の用言が含まれている場合、後ろのひらがな用言は、文末の定型表現の一部である可能性も大きい。

このことを調査するために、形態素列 M に複数の P が含まれている場合、後ろの P を F と書き換えると、新たにどのような文末機能表現候補が抽出されるかを調べた。その結果を表 3 に示す。この表には、頻度の高い表現を示したが、これらの大半は、文末機能表現候補として検討する値する。しかしながら、頻度の低い表現の中には多数のゴミが含まれる。そのため、単純に P を F に書き換えるのではなく、もう少し洗練されたヒューリスティックを考案する必要がある。

今後、このような改善を進めるとともに、得られた候補リストと見出し語リストの照合や、候補リストを用いた辞書のカバレジの推定等を行なっていく予定である。

謝辞　本研究では、CD-毎日新聞データ集 (2005 年版) を使用した。

参考文献

[1] 桝田達也, 佐藤理史. 文末表現シソーラスの設計と編纂. 言語処理学会 第 16 回年次大会 発表論文集, 2010.

[2] 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.

表 2: 抽出結果

文末機能表現	頻度 J	頻度 U	頻度 I
V. ことができるのではないでしょうか	3	3	3
V. ないようにしなければならない	3	3	3
V. てもいいのではないでしょうか	–	4	4
V. なければならないのでしょうか	4	4	4
V. なければならなくなりました	3	3	3
V. なければならなくなっている	3	3	3
V. もいいのではないでしょうか	4	–	–
V. ていたのではないでしょうか	–	3	3
V. なければならないものとする	3	3	3
V. ているのではないでしょうか	–	13	12
V. ことができるようになった	16	16	16
V. ているわけではありません	–	6	5
V. ているのではないだろうか	–	13	12
V. られるようになったという	3	3	3
V. ないのではないでしょうか	4	5	6
V. いたのではないでしょうか	3	–	–
V. れていなかったとしている	2	3	3
V. もいいのではないだろうか	4	–	–
V. てくれるようになりました	–	3	3
V. てこなかったのではないか	–	3	3
V. ていなかったのではないか	–	4	4
V. うとしているのではないか	–	–	3
V. ていかなくてはなりません	–	3	3
V. ていかなければなりません	–	9	9
V. れることになってしまった	4	4	4
V. いるのではないでしょうか	12	–	–
V. ないようにしているという	3	3	3
V. たのではないかとみられる	–	3	3
V. たのではないかとみている	–	6	6
V. なければならないのだろう	3	2	3
V. なければならないからだ	6	6	6
V. なければならないだろう	6	6	6
V. いかなければなりません	8	–	–
V. いるわけではありません	6	–	–
V. いかなくてはなりません	3	–	–
V. なければならないはずだ	3	3	3
V. たのではないでしょうか	–	9	9
V. なければならなくなった	12	12	12
V. れるようになったという	5	5	5
V. こなかったのではないか	3	–	–
V. ていたということだろう	–	3	3
Ai. ことではないでしょうか	5	–	–

J, U, I は、JUMAN, UniDic, IPAdic を表す。
V は動詞、Ai は形容詞を表す。

表 3: 新たに抽出される表現 (JUMAN の場合)

文末機能表現	頻度
V. にとどまった	154
V. といえる	125
V. のかもしれない	115
V. にとどめた	92
Ai. かもしれない	77
V. はいけない	67
V. わけにはいかない	62
Ai. があった	55
Ai. さをにじませた	48
N. のかもしれない	47
V. ざるをえない	43
Ai. のかもしれない	43