機能表現を考慮した日本語係り受け解析器学習のためのコーパス作成

土屋雅稔[†1] 注連隆夫[†2] 宇津呂武仁[†4]
松吉俊[†2,†3] 佐藤理史[†3] 中川聖一[†5]

1. はじめに

機能表現とは,「にあたって」や「をめぐって」のように，2 つ以上の語から構成され，全体として 1 つの機能的な意味をもつ表現である．一方，この機能表現に対して，それと同一表記をとり，内容的な意味をもつ表現が存在することがある．例えば，文 (1) と文 (2) には,「にあたって」という表記の表現が共通して現れている．

(1) 出発するにあたって，荷物をチェックした
(2) ボールは，壁にあたって跳ね返った

文 (1) では，下線部はひとかたまりとなって,「機会が来たのに当面して」という機能的な意味で用いられている．それに対して，文 (2) では，下線部に含まれている動詞「あたる」は，動詞「あたる」本来の内容的な意味で用いられている．これらの表現においては，機能的に用いられている場合と，内容的に用いられている場合とを識別する必要がある．以下，文 (1) , (2) の下線部のように，表記のみに基づいて判断すると，機能的に用いられている可能性がある部分を**機能表現候補**と呼ぶ．

日本語複合辞用例データベース[1](以下，**用例データベース**と呼ぶ) は，機能表現の機械処理を研究するための基礎データを提供することを目的として設計・編纂されたデータベースである．この用例データベースは，現代語複合辞用例集[2](以下，**用例集**と呼ぶ) に収録されている 125 種類の複合辞および，その異形 (合計 337 種類の機能表現) を対象として，機能表現候補と一致する表記のリストと，個々の機能表現候補に対して最大 50 個の用例を毎日新聞 (1995 年) から収集したものを収録している．そして，各機能表現候補が文中において果たしている働きを表す 6 種類の判定ラベル (**表 1**) を人手で付与している．

用例データベースによると，機能的用法か内容的用法かの判別が必要な表現は，少なくとも 111 表現存在する．これらの表現に対する，既存の解析系の扱いを調べてみた．形態素解析器 JUMAN☆1と構文解析器 KNP☆2の組み合わせによって，機能的な意味で用いられている場合と内容的な意味で用いられている場合とが識別される可能性がある表現は 111 表現中 43 表現である．また，形態素解析器 ChaSen☆3と構文解析器 CaboCha☆4の組合わせを用いた場合には，識別される可能性がある表現は 111 表現中 40 表現である．

以上より，機能表現は，日本語の文構造を把握する時に重要な役割を果たしているにも関わらず，従来の自然言語処理における取り扱いは不十分であることがわかる．このような現状を改善するには，機能表現候補の用法を正しく識別する検出器と検出器によって検出される機能表現を考慮した係り受け解析器が必要である．本論文では，そのような検出器と係り受け解析器を機械学習的手法によって作成する場合の訓練データとして用いることができるコーパスを作成する方法を述べる．

本論文の構成は以下の通りである．最初に，機能表現検出器と機能表現を考慮した係り受け解析器の構成について述べ (2 章)，次に，コーパスを作成する手順を示す (3 章)．更に，コーパスの内容を分析した上で (4 章)，結論を述べる (5 章)．

2. 機能表現を考慮した係り受け解析

機能表現検出器と機能表現を考慮した係り受け解析器を実現するには，幾つかの構成法が考えられるが，本論文では**図 1** のような構成を考える．まず，ChaSen によって形態素解析を行う．次に，形態素解析結果に対して機能表現検出器を適用し，機能表現を検出する．その際，機能表現を構成している形態素列を連結し，機能表現は 1 つの形態素として出力する．最後に，その出力結果に対して，機能表現を考慮した係り受け解析器を用いて，係り受け解析を行う．

機能表現を考慮した係り受け解析器としては，人手規則による解析器と，機械学習による解析器の 2 通りが考えられる．しかし，人手規則による解析器を実現するには，通常の係り受け解析規則と整合性を取りながら，機

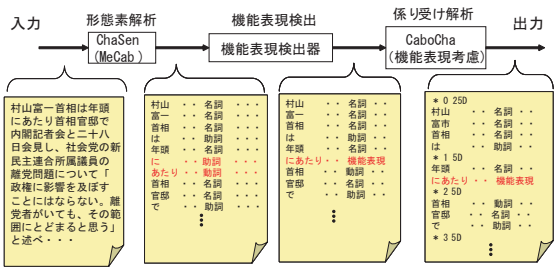

図 1 機能表現を考慮した係り受け解析

†1 豊橋技術科学大学 情報メディア基盤センター
†2 京都大学大学院 情報学研究科
†3 名古屋大学大学院 工学研究科
†4 筑波大学大学院 システム情報工学研究科
†5 豊橋技術科学大学 工学部情報工学系

☆1 http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
☆2 http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html
☆3 http://chasen.naist.jp/hiki/ChaSen/
☆4 http://chasen.org/~taku/software/cabocha/

表 1　判定ラベル体系

判定ラベル	判定単位	読み	内容 vs 機能	用法	例文
B	不適切				(3) 不平等条約を盾にとり、ゆすりに等しい権利を主張している。
Y	適切	不一致			(4) 法律上は困難でも，もう少し組織的に救援活動に参加する ···
C	適切	一致	内容的	内容的用法	(5) まな板にとってていねいに納豆のタタキを作りみそ汁の実にする ···
F	適切	一致	機能的	用例集で説明されている用法	(6) 受験などでは倍率が上がったところで入学金があがることはない.
A	適切	一致	機能的	接続詞的用法	(7) ところで，全国の桜の名所では近年，樹勢の衰えが目立ち，···
M	適切	一致	機能的	その他の機能的用法	(8) 浜ノ島はあと一歩のところで勝ち星に結び付かず負け越した.

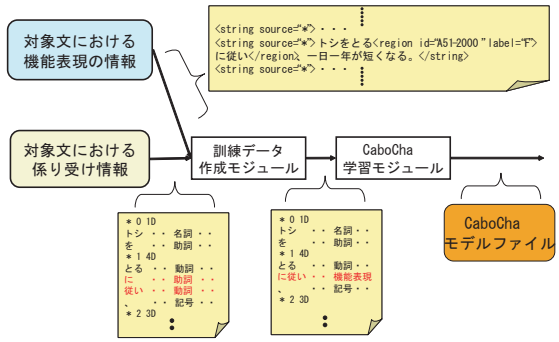

図 2　機能表現を考慮した係り受け解析器の学習

能表現を考慮した係り受け解析規則を整備していく必要があり，非常な困難を伴うことが予想される．そのため，本論文では，機械学習による解析器を考える．具体的には，SVM を用いた統計的係り受け解析手法の学習・解析ツールとして CaboCha を利用し，CaboCha の係り受け解析モデルの訓練データを機能表現を考慮したデータに変換することにより，機能表現を考慮した係り受け解析器を実現する．

3. コーパスの作成

機能表現を考慮した係り受け解析器の訓練データを作成するには，2 つの情報が必要である (**図 2**)．第 1 は，対象とする文における機能表現の情報であり，第 2 は，対象とする文における係り受けの情報である．

3.1 機能表現候補に対する判定ラベルの付与

機能表現の情報を作成するには，対象とする機能表現のリスト，および，機能表現候補が文中において果たしている働きを表現する方法という仕様にあたる部分と，その仕様に合致する情報を作成する手順の 2 つを決める必要がある．

3.1.1 対象とする機能表現と判定ラベル

本コーパスの作成にあたっては，用例データベースに収録されている 337 種類の機能表現を対象とする．これらの機能表現は，用例集を参照すると，助詞型・接続辞類，助詞型・連用辞類，助詞型・連体辞類，助動詞型，接続詞型の 5 種類に分類できる．

各々の機能表現候補が文中において果たしている働きを表す方法としては，機能表現候補が文中でどのような働きをしているかを表す判定ラベルを，機能表現候補に対して付与するという方法を取る．判定ラベルには，用例データベースと同じ 6 種類の判定ラベル (表 1) を用い，用例データベースと同一の判定基準で判定を行う．

3.1.2 機能表現候補の列挙と判定ラベルの付与

本コーパスの作成にあたっては，対象とする文に出現する全ての機能表現候補に対して，判定ラベルを付与する必要がある．そのため，最初に，対象となる文に出現する全ての機能表現候補を機械的に列挙し，次に，それらの機能表現候補に対して判定ラベルを人手で付与するという 2 段階の処理を行う．

全ての機能表現候補を列挙するため，2 通りの手法を適用する．第 1 に，機能表現候補と一致する表記が出現し，かつ，表記と一致した部分の先頭と末尾が形態素境界となっている箇所を収集する．第 2 に，文中の活用語を 1 つずつ基本形に置き換えた文を生成し，その文に対して機能表現候補と一致する表記が含まれているか調べて，表記の末尾形態素が活用して用いられている場合を収集する．

次に，列挙された全ての機能表現候補に対して，判定ラベルを付与する作業を人手により行う．ただし，複数の機能表現候補が部分的に重複して出現している場合には，1 つの機能表現候補に対して判定ラベルを付与すると，残りの機能表現候補に対する判定作業は省略できる場合がある．例えば，文 (9) には,「〜にしても (A34–1000)」と「〜てもいい (B30–1000)」の 2 つの機能表現候補が部分的に重複して現れている．

(9)　虫食いだらけ にしてもいいのだろうか.

この時，機能表現候補「てもいい」に判定ラベル F を付与すると，機能表現候補「にしても」に対する判定ラベルは自動的に B に決まるので，機能表現候補「にしても」に対する判定作業は省略することができる．

判定ラベル付与を円滑に行うため，**図 3** のようにウェブブラウザ上で動作する作業環境を作成した．この作業環境は，機能表現候補に対して可能な判定ラベルの一覧をプルダウンメニューの形式で表示し，容易に判定ラベルを付与できるように設計されている．また，複数の機能表現候補が部分的に重複している場合には，それらの候補を一括して扱い，必要最小限の作業のみで判定が完了できるようになっている．

3.2 係り受け情報の付与

日本語の係り受け情報としては，京都テキストコーパス[3] が広く利用されている．仮に，京都テキストコーパ

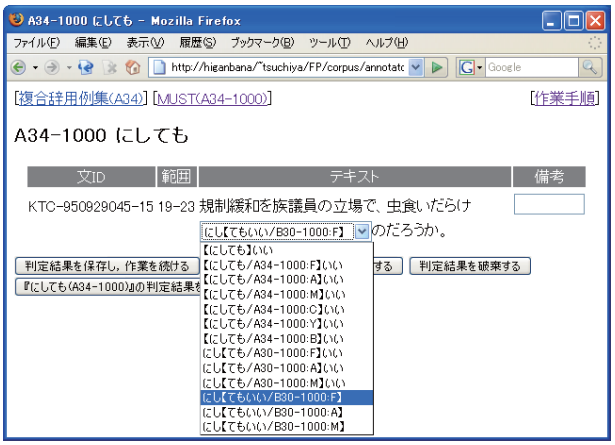

図 3　機能表現候補に対する判定ラベル付与インターフェース

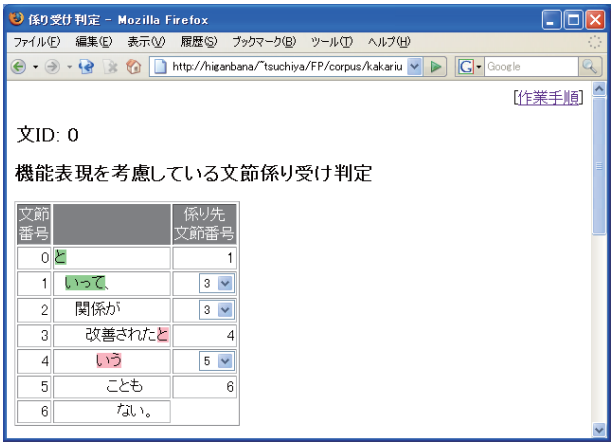

図 4　係り受け情報付与インターフェース

スに現れる全ての機能表現候補に判定ラベル付与を行い，そのコーパスだけを用いて，機能表現を考慮した係り受け解析器の訓練と評価が可能ならば，新たな係り受け情報は不要である．しかし実際には，京都テキストコーパスでは，種々の機能表現候補をバランス良く集めることは考慮されていないので，一部の機能表現については，訓練と評価が正しく行えない．

そのため，訓練データおよび評価データとして，適切な文を新たに補充する必要がある．訓練データとして用いる文には，文中の全ての係り受け関係が必要であるが，評価データとして用いる文には，評価対象となっている機能表現を含む文節が関係する係り受け関係のみが必要である．**図 4** は，評価用データを作成するためのウェブブラウザ上で動作する作業環境である．評価対象となっている機能表現を含む文節の係り先と，その文節に対して係っている文節を変更することだけができるよう設計されており，作業者は，必要最小限の操作で評価データを作成できる．

3.3　機能表現情報と係り受け情報の統合

機能表現情報と係り受け情報を統合して，機能表現を考慮した係り受け解析器の訓練データを作成する手順の概略は，以下の通りである．

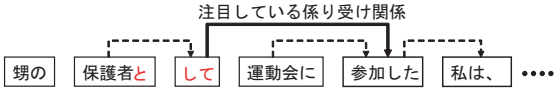

図 5　係り受け解析例 (機能表現を考慮しない場合)

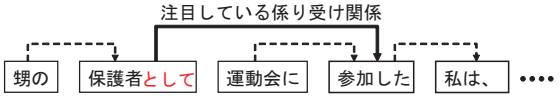

図 6　係り受け解析例 (機能表現を考慮した場合)

表 2　機能表現候補数

機能表現候補数 n	小項目数	例
$1000 \leq n$	3 (1%)	～という (A82–1000)
$500 \leq n < 1000$	7 (2%)	～によると (A61–2000)
$100 \leq n < 500$	34 (10%)	～つつある (B35–1000)
$50 \leq n < 100$	28 (8%)	～をめぐって (A73–1000)
$10 \leq n < 50$	68 (20%)	～一方だ (B5–1000)
$5 \leq n < 10$	31 (9%)	～と思えば (A9–4000)
$1 \leq n < 5$	55 (16%)	～ては駄目だ (B29–6000)
$n = 0$	111 (33%)	～に関しまして (A46–1100)
計	337	

表 3　判定ラベル数

判定ラベル						
F	A	M	C	Y	B	計
14362	215	3686	1645	1733	186	21827

最初に，**図 5** のように機能表現を考慮していない係り受け構造を用意する．次に，それに含まれている機能表現に対して，機能表現を構成している形態素列を連結し，各々の機能表現を 1 つの形態素と見なすようにする．同時に，文節の分割位置や係り先を調整する．このような処理を行うと，**図 6** のように機能表現を考慮した係り受け構造が得られる．

詳細な手順については，注連ら[4] を参照されたい．

4.　コーパスの分析

京都テキストコーパスに収録されている文を対象として，全ての機能表現候補の列挙を行った結果を**表 2** に示す．対象とする 337 表現の内，111 表現は，機能表現候補としても京都テキストコーパスには現れなかった．用例データベースでは，機能表現候補としてもまったく現れなかった表現は 33 種類であり，種々の機能表現候補をバランス良く集めることを考慮したか否かの違いが現れている．**表 3** に，京都テキストコーパス中に出現した 21827 箇所の機能表現候補を判定した結果を示す．なお，判定ラベル付与インターフェースの動作記録によると，実際に判定作業を行った機能表現候補は 20095 箇所であり，機能表現候補が部分的に重複している場合の判定作業を効率化できたことが分かる．

判定ラベル F は，その機能表現候補が，用例集で説明された用法で用いられていることを示す．用例集には，用法に関する説明文と例文が収録されているため，判定ラベル F には，他の判定ラベルに比べて，判定時に比較的しっかりとした根拠があるという特徴がある．判定ラベ

表 4 判定ラベル F の出現率

出現率 x	小項目数	例
x= 100%	99 (43.8%)	～うものなら (A11–1000)
95% <x< 100%	16 (7.1%)	～によると (A61–2000)
5% ≤x≤ 95%	86 (38.1%)	～にしても (A34–1000)
x< 5%	25 (11.1%)	～と思うと (A9–2000)
計	226	

$$x = \frac{\text{判定ラベル F が付与された機能表現候補数}}{\text{機能表現候補数}}$$

表 5 判定ラベル F の出現頻度

出現頻度 f	小項目数	例
1000 ≤ f	2 (1%)	～として (A62–1000)
500 ≤ f < 1000	5 (2%)	～について (A53–1000)
100 ≤ f < 500	25 (11%)	～にとって (A56–1000)
50 ≤ f < 100	18 (8%)	～ばかりだ (B18–1000)
10 ≤ f < 50	65 (29%)	～において (A39–1000)
5 ≤ f < 10	32 (14%)	～に応じた (A42–1011)
1 ≤ f < 5	57 (25%)	～ても仕様がない (B31–1000)
f = 0	22 (10%)	～を問わずに (A68–2000)
計	226	

f = 判定ラベル F が付与された機能表現候補数

ル F が付与された機能表現候補の出現率を**表 4** に，出現頻度を**表 5** に示す．表 4 より，99 表現は，単純に表記との文字列一致を行うだけで，用例集の用法で用いられている機能表現であると分かる．また，判定ラベル F の出現率が 5%未満となっている表現は，25 種類ある．それに対して，用例データベースでは，接続制約を加えた補充収集を行っているので，判定ラベル F の出現率が 5%未満の表現は 3 種類とより少なくなっている．

判定ラベル F,A,M は，その機能表現候補が，何らかの機能的な働きをしているという点で共通する判定ラベルである．つまり，**表 6** は機能的用法で用いられている機能表現候補の出現率を，**表 7** は出現頻度を示している．表 6 より，82 種類の機能表現は，京都テキストコーパスの範囲に限ってみても，機能的用法の出現率が 5%以上かつ 95%以下となっており，用法の曖昧性が存在している．用例データベースによると，この 82 種類の表現の内，毎日新聞 (1995 年) に 50 回以上の頻度で機能表現候補が出現した表現は 62 種類である．これらの表現は，出現頻度の観点から見ても，また，用法の曖昧性の観点から見ても，最初に取り組むべき重要な表現と考えられる．

このようなコーパスの作成にあたっては，作業者間の判定の揺れが大きな問題となってくる．中でも，複数の機能表現候補が部分的に重複して現れている場合は，特に判定が揺れやすいと考えられる．そのため，そのような候補から 517 個を選び，2 人の作業者により独立に判定作業を行い，判定結果が実際に一致した割合 P_a と，一致度 κ を求めた．その結果と，用例データベースにおける P_a と κ を，**表 8** に示す．表 8 より，特に判定が揺れやすいと考えられる候補に対しても，用例データベースと殆んど変わらない一致度が達成され，本コーパスの判定結果はおおむね信頼できることが分かる．

表 6 判定ラベル F,A,M の出現率

出現率 x'	小項目数	例
x'= 100%	111 (49.1%)	～ごとに (A23–1000)
95% <x'< 100%	19 (8.4%)	～としては (A63–1000)
5% ≤x'≤ 95%	82 (36.3%)	～にしても (A34–1000)
x'< 5%	14 (6.2%)	～といけない (B29–2000)
計	226	

$$x' = \frac{\text{判定ラベル F,A,M が付与された機能表現候補数}}{\text{機能表現候補数}}$$

表 7 判定ラベル F,A,M の出現頻度

出現頻度 f'	小項目数	例
1000 ≤ f'	3 (1%)	～との (A82–2000)
500 ≤ f' < 1000	6 (3%)	～ものだ (B1–1000)
100 ≤ f' < 500	29 (13%)	～ところが (A21–1000)
50 ≤ f' < 100	24 (11%)	～こともある (B12–3000)
10 ≤ f' < 50	66 (29%)	～までだ (B16–1000)
5 ≤ f' < 10	29 (13%)	～ものなら (A28–1000)
1 ≤ f' < 5	56 (25%)	～ても仕様がない (B31–1000)
f' = 0	13 (6%)	～かと思うと (A9–1000)
計	226	

f' = 判定ラベル F,A,M が付与された機能表現候補数

表 8 作業者間の判定の一致度

区別する判定ラベル		京都テキストコーパス		用例データベース	
X	$\bar{X}$	P_a	κ	P_a	κ
B	Y, C, F, A, M	0.98	0.79	0.97	0.77
F, A, M	B, Y, C	0.97	0.87	0.93	0.73
F	B, Y, C, A, M	0.96	0.79	0.96	0.85

5. おわりに

本論文では，機能表現検出器と機能表現を考慮した係り受け解析器の訓練データとして利用できるコーパスを作成する方法を提案した．そのようなコーパスを作成するには，対象とする文の係り受け情報と，対象とする文に現れる全ての機能表現候補に判定ラベルが付与されていることが必要である．実際に，必要な情報を整備するための作業環境を作成し，京都テキストコーパスに現れる全ての機能表現候補に判定ラベルを付与した．その結果を用いて，作成した作業環境の有効性と，京都テキストコーパスと用例データベースの違いを示した．このコーパスを利用すると，機能表現検出器と機能表現を考慮した係り受け解析器を作成することができ，機能表現を考慮していない従来の解析系と比較して，機能表現検出および係り受け解析の精度が改善した[4]．

参考文献

1) 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一: 日本語複合辞用例データベースの作成と分析, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741 (2006).
2) 国立国語研究所: 現代語複合辞用例集 (2001).
3) 黒橋禎夫, 長尾眞: 京都大学テキストコーパス・プロジェクト, 言語処理学会第 3 回年次大会発表論文集, pp. 115–118 (1997).
4) 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史: 機械学習を用いた日本語機能表現のチャンキングおよび係り受け解析, 言語処理学会第 12 回年次大会論文集 (2007). B3–4.