語彙化文法における語彙項目の構造的特徴に基づく自動分類

大内田賢太† 吉永直樹* 二宮崇‡* 宮尾祐介* 辻井潤一‡*
† 東京大学 理学部情報科学科 ‡CREST, 科学技術振興事業団
* 東京大学大学院 情報理工学系研究科 コンピュータ科学専攻
{oouchida, yoshinag, ninomi, yusuke, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

語彙化文法 [1] の流れを汲む HPSG (主辞駆動句構造文法: [2]) や LTAG (語彙化木接合文法: [1]) などの文法枠組は，文に項構造などの豊かな意味表現を与えることができるため，高度な自然言語処理アプリケーションに利用されることが期待されている．しかしながら，語彙化文法の性質上，単語に割り当てられる語彙項目の数が文法規則の数に比して圧倒的に多くなる．従って，語彙項目を人手で互いに矛盾無く記述することが困難なため，広範で一貫した文法は現状では存在していない．

このような膨大な語彙項目を統一的に扱うために，多くの語彙化文法の枠組では，(i) 品詞および下位範疇化フレーム (形容詞、他動詞や二重他動詞) と (ii) 構文構造 (受動文・命令文など) の組み合わせにより表現することで，膨大な語彙項目を体系的・効率的に表現する試みが行われてきた [3, 4, 5, 6]．本論文では前者に基づく分類を**統語的クラス**，後者に基づく分類を**構造的クラス**と呼ぶ．語彙項目の体系化は，人手による文法開発のうち最も労力を必要とする部分の一つであり，語彙化文法を開発する際の障害となっている．また同時に，言語学的な一般化を捉えるという意味で言語学の分野でも重要な問題となっている．

本研究では，このような語彙化文法の語彙項目の体系化を自動化するために，統語的クラスに分類された語彙項目を構造的クラスへ自動分類する手法を提案する．本研究により，文法に暗黙のうちに含まれている言語学的な一般性を自動的に抽出し，人手で語彙項目を体系化する労力を大幅に削減することを目指す．さらには，最近の大規模な語彙化文法で特に注目されている括弧つきコーパスからの自動獲得により獲得された文法 [7, 8, 9] に，既存の統語的クラスへの自動分類手法 [10] と本手法を段階的に適用することで，人手によりメンテナンス可能な体系化された文法を最小の労力で獲得することが出来るようになる．

我々は，本手法を既存の人手で書かれた大規模語彙化文法である XTAG 英文法に対して適用し，得られた構造的クラスを人手による構造的クラスと比較してその一致度を調べることで本手法の有効性を評価した．

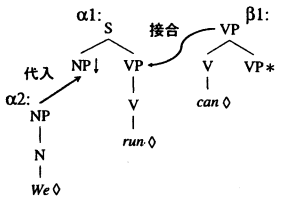

図 1: LTAG の語彙項目と文法規則

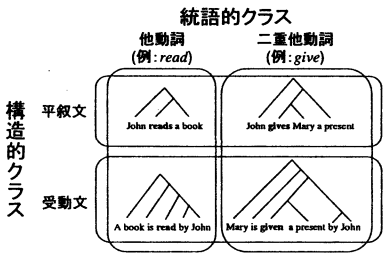

図 2: 語彙化文法の語彙項目の体系化

2 語彙化文法の体系化

語彙化文法 [1] は，(i) 単語特有の語彙的・統語的な制約を記述した**語彙項目**と，(ii) 言語の一般的な文法構造を記述する**文法規則**とから構成される文法枠組のクラスである．図 1 に LTAG の語彙項目である木構造 ($\alpha 1$, $\alpha 2$, $\beta 1$) とそれらを組み合わせる文法規則である代入と接合を示す．語彙項目はこの例のように，単語の品詞および下位範疇化フレームと，単語がとりうる構文構造に対する制約が記述される．

語彙化文法に属する多くの文法枠組では言語学的な一般化および効率的な語彙項目の表現を目的として，語彙項目の体系化が試みられている [3, 4, 5, 6]．語彙項目の体系化とは，具体的には語彙項目を品詞および下位範疇化フレームに基づき分類し (統語的クラスの分類)，語彙項目が表す構文構造に従い分類する (構造的クラスの分類) ことである．図 2 に read および give の語彙項目の分類を示す．統語的クラスについては，同じ下位範疇化フレームを持つという明確な分類基準がある．一方，

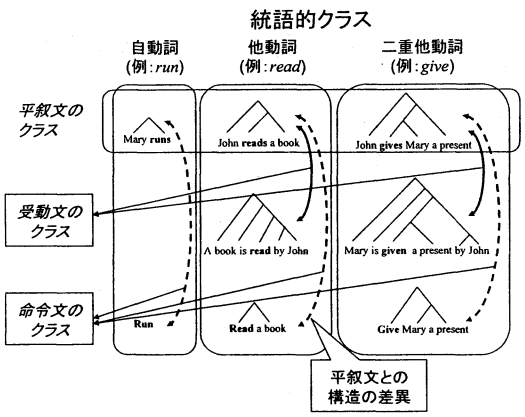

図 3: 語彙化文法の語彙項目の体系化

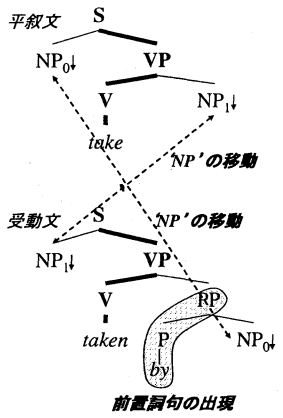

図 4: 受動文の語彙項目の構造的特徴

構造的クラスについては，下位範疇化フレームなどの語彙項目に含まれる情報だけでなく，平叙文の語彙項目との間の統語的変形関係も重要な分類基準となる．

本研究では，入力の語彙項目の集合は，1) 統語的クラスに基づき分類されており，2) 構造的クラスについては平叙文のクラスに対してのみ既に分類されていることを前提として，与えられた語彙項目の集合を構造的クラスに分類することを目標とする．このうち，1) については既存の研究が存在する [10] が，2) については，今後の研究課題になると思われる．このような前提の下で，各語彙項目は平叙文の語彙項目から統語的変形によりに暗黙のうちに生成されていると考え，この統語的変形を語彙項目の構造の差異という形で得て，その差異に基づき分類を行う（図 3）．

3 構造的素性

前節で述べたように，本研究では構文構造の特徴を，同じ統語的クラスに属する平叙文の語彙項目との構造的差異として捉える．図 4 は LTAG において受動文の語彙項目と平叙文の語彙項目との構造的差異を図示したものである．図中で斜線で示された構造は受動文の語彙項目に新たに現れた構造であり，一方，点線で示された対応は下位範疇化要素の順列の変化を表している．我々は語彙項目と平叙文の語彙項目との間の構造的差異を，与えられた語彙項目の**構造的素性**として導入する．一般的に，語彙項目の構造的素性としては，以下の 2 種類があると考えられる．

出現素性 平叙文の語彙項目に対して消滅する下位範疇化要素と付加的に出現する下位範疇化要素のペア．例えば受動文の語彙項目の場合は “by” 前置詞句が付加的に出現し，命令文の語彙項目の場合は主語位置の名詞句が消滅する．

移動素性 平叙文の語彙項目に対する下位範疇化要素の下位範疇化の順列の変化．例えば受動文や目的語の wh 移動の語彙項目の場合は，平叙文で目的語，主語の順で行われる下位範疇化が逆転する．本研究では名詞句の下位範疇化の順列の移動についてのみ考える．

本研究では，構造的特徴として移動素性を抽出するために，入力の語彙項目の集合に対して，2 節で述べた 2 つの前提に加えて，3) 語彙項目の下位範疇化要素に意味的役割が与えられており，同じ統語的クラスの二つの語彙項目を比較したときに下位範疇化要素の対応関係がとれているということも前提とする．

4 構造的素性の抽出手法

本節では，前節で導入した出現素性と移動素性を LTAG の枠組で抽出する手法について述べる．

アルゴリズム

2 節および前節であげた 1), 2), 3) の前提を満たす語彙項目の集合の各要素を入力とし，出力として前節で述べた出現素性および移動素性を出力するアルゴリズムについて述べる．

図 5 に LTAG における出現素性の抽出アルゴリズムを示す．`extract_appearence_feature` は，語彙項目 T を受け取り，T と同じ統語的クラスの平叙文の語彙項目 T_b とを比較し，出現素性を返す関数である．LTAG においては 2 節でみたように，語彙項目が木構造で表現され，木構造中のノードにより下位範疇化要素に対する制約が表現される．そこで，まず，入力の語彙項目を表す木構造中のノードと，平叙文の語彙項目を表す木構造中のノードとの差分をとる（図 5 中 (1)）．次に，その差分のノードのみから構成される部分構造を作り，できた部分構造と木構造中で繋がるノードからなる部分構造を作る（図 5 中 (2)）．このようにして，新たに出現する

```
procedure extract_appearance_feature(T)
begin
  T_b := declarative(T)
  F_1 := extract_appearance_feature_sub(T, T_b)
  F_2 := extract_appearance_feature_sub(T_b, T)
  return (F_1, F_2)
end

procedure extract_appearance_feature_sub(T1, T2)
begin
  N := node_list(T1) − node_list(T2)                    ... (1)
  F := φ
  while (N ≠ φ)
    n := get_node(N)
    f := create_partical_tree(n, N, T1)                 ... (2)
    N := N − node_list(f)
    F := F ∪ {f}
  end while
return F
end

declarative:入力と同じ統語的クラスの平叙文の語彙項目を返す
node_list:  木構造を受け取り，そのノードのリストを返す
get_node:   リストの要素を一つ返す
create_partial_tree:
      第三引数の語彙項目の部分構造のうち，第一引数のノードを含み，
      第二引数のノードのみからなる最大の部分構造を得て，その部分
      構造に繋がるノードを含めた部分構造を返す
```

図 5: LTAG における出現素性の抽出アルゴリズム

```
procedure extract_movement_feature(T)
begin
  T_b := declarative(T)
  F := φ
  I := make_index(T, T_b)                               ... (1)
  foreach i (I)
    if ( get_parent(i, T) = get_parent(i, T_b) )
      I := I ∪ {i}                                      ... (2)
    end if
  end foreach
  return F
end

declarative:入力と同じ統語的クラスの平叙文の語彙項目を返す
make_index: 木構造に与えられている意味役割に従って、二つの木構造
            で同じ意味役割の NP の葉ノードに同じインデクスを与え，
            最終的に与えたインデクスの集合を返す
get_np_nodes:木構造の NP ノードを返す
get_parent:  インデクスを与えられた NP ノードの親ノードを返す
```

図 6: LTAG における移動の抽出アルゴリズム

部分構造のリストと消滅する部分構造のリストを作り，これらのペアを出現素性として返す．

次に，図 6 に移動素性の抽出アルゴリズムを示す．**extract_movement_feature** は，語彙項目 T を受け取り，T と同じ統語的クラスに含まれる平叙文の語彙項目 T_b とを比較し，移動素性を返す関数である．LTAG においては下位範疇化要素を下位範疇化する順列は木構造中の葉ノードの位置により表現される．そこで，まず，T と T_b の意味的制約により対応がとれている NP ノードに対して，同じインデクスをつけていく（図 6 中 (1)）．次に，同じインデクスをつけられた各 NP ノードのペアに対し，それぞれの親のノードのラベルが同じでなければ，そのインデクスをリストに入れ（図 6 中 (2)），最終的に得られたリストを移動素性として返す．

これらのアルゴリズムを他動詞と二重他動詞の語彙項目に適用することで得られる受動文の構造的素性を図 7 および図 8 にあげる．両者から同じ構造的素性を抽出

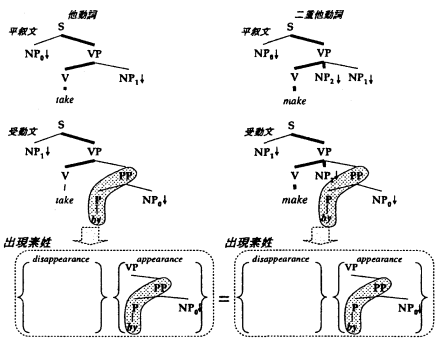

図 7: LTAG の受動文の構文構造を表す語彙項目の出現素性

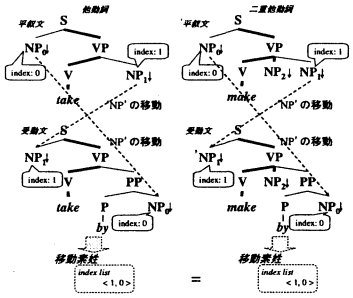

図 8: LTAG の受動文の構文構造を表す語彙項目の移動素性

できることを示している．

5 実験

本手法を語彙化文法の一つである LTAG [1] を対象として実装した．入力として大規模 LTAG 文法である XTAG 英文法 [11] の tree family と呼ばれる動詞の統語的クラス 57 個とそれに属する 1,008 個の語彙項目テンプレートを入力として用いた．その結果，63 種類の出現素性，5 種類の移動素性を抽出し，134 個のクラスを得た．

次に，得られた 134 のクラスと XTAG 英文法の 74 の構造的クラスとを比較することで，本手法の有効性と人手による体系化の差異を考察する．我々はまず、人手による平叙文以外の 73 の構文的クラスについて、本手法で得られたクラスの中に一致するものがあるかどうかを調べた．その結果を表 1 に示す．人手による構文的クラスのうち 34(46.6%) の構文的クラスは得られたクラスの中に一致するものがあり（表 1 中 $\mathcal{A}$），31(42.5%) の構文的クラスについては得られたクラスの中に細分化されたクラスがあった（表 1 中 $\mathcal{B}$）．残りの 8(10.9%) については得られたクラスの中では一部が他のクラスと混ざっていた（表 1 中 $\mathcal{C}$）．この結果から，我々の手法によりある程度人の直感に近いあるいはより詳細な構造的クラス

表 1: 人手による構造的クラスの分析: ($\mathcal{A}$) 得られたクラスの中に一致するものがあった構造的クラス, ($\mathcal{B}$) 得られたクラスの中に細分化されたクラスがあった構造的クラス, ($\mathcal{C}$) 得られたクラスの中では一部が他のクラスと混ざっていた構造的クラス

人手による構造的クラスのグループ	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	計
wh 移動	2	4	0	6
wh 移動+受動文	8	0	2	10
関係詞節	0	7	2	9
関係詞節+受動文	16	4	2	22
受動文	5	7	0	12
形容詞節	1	0	0	1
名詞にかかる他動詞の受動文	1	0	0	1
動名詞	0	7	0	7
動名詞+受動文	0	4	0	4
能格動詞の名詞化	1	0	0	1
Y/N 疑問文	0	1	0	1
挿入句	0	1	0	1
修飾詞	0	1	0	1
補語の順序が交替した平叙文	0	2	0	2
命令文	0	1	0	1
話題化	0	0	1	1
名詞節	0	0	1	1
計	34	31	8	73

が得られているといえよう. 次節で我々は上記の $\mathcal{B}$, $\mathcal{C}$ に分類された構文的クラスについてそれぞれ考察する.

6 考察

6.1 より細かい分類を得た構文的クラス

人手による構文的クラスのうち, (表 1 中 $\mathcal{B}$) に含まれる構文的クラスは我々の手法によりより細かい部分クラスに分類された. これらの細かい部分クラスの語彙項目の構造的素性には, 多少なりと, 統語的クラスの影響が見られた. 例えば図 9 は, 人手による構造的クラスでは一つにまとまる語彙項目が異なった構造的素性を持っていた例である. 人手による構造的クラスと同一のクラスを得ることとを目指すならば, このような構造的素性に現れる統語的クラスの影響を吸収するために, 似通った構造的素性を持つ語彙項目を一つの構造的クラスにまとめるなどの方法が考えられよう.

6.2 他の構造的クラスと混ざった構造的クラス

今回抽出した構造的素性だけでは分けることができない語彙項目がわずかながら (10.9%) 存在した (表 1 中の $\mathcal{C}$). この中には含まれる構文的クラスのうち "wh 移動" と "話題化" については, 下位範疇化要素の移動と出現の差異に着目するとこの 2 つの構文構造は区別できないことが分かった. この場合, 人手による構造的クラスと同一のクラスを得るためには, 語彙項目中の他の情報を素性として抽出する必要があろう.

7 まとめ

本稿では, 語彙化文法の語彙項目を各語彙項目の構造的素性に従い分類する手法を提案した. 我々は本手法を既

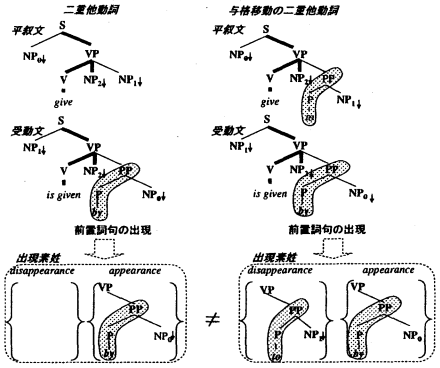

図 9: 統語的クラスの影響により一つのクラスにならなかった人手による構文的クラス

存の人手で書かれた大規模語彙化文法である XTAG 英文法に対して適用し, 1,008 の動詞の語彙項目を 134 のクラスに分類することに成功した. 得られたクラスと人手で書かれたクラスを比較した結果, 人手による構造的クラスの 89.1%について, その構造的グラス自体か, その構造的クラスを細分類したクラスを得ることができた.

参考文献

[1] Y. Schabes, A. Abeillé, and A. K. Joshi. Parsing strategies with 'lexicalized' grammars: Application to Tree Adjoining Grammars. In *Proc. of the 12th COLING*, pages 578–583, 1988.

[2] C. Pollard and I. A. Sag. *Head-Driven Phrase Structure Grammar*. CSLI Publications, 1994.

[3] D. Flickinger. *Lexical Rules in the Hierarchical Lexicon*. PhD thesis, Stanford University, 1985.

[4] B. Carpenter. The generative power of Categorial Grammars and Head-Driven Phrase Structure Grammars with lexical rules. *Computational Linguistics*, 3(17):301–313, 1991.

[5] T. Becker. Patterns in metarules. In *Proc. of TAG+3*, pages 9–11, 1994.

[6] M.H. Candito. A principle-based hierarchical representation of LTAGs. In *Proc. of the 16th COLING*, pages 578–583, 1996.

[7] F. Xia. Extracting Tree Adjoining Grammars from bracketed corpora. In *Proc of the fifth NLPRS*, pages 398–403, 1999.

[8] J. Hockenmaier and M. Steedman. Generative models for statistical parsing with combinatory categorial grammar. In *Proc. of the 40th ACL*, pages 335–342, 2002.

[9] Y. Miyao, T. Ninomiya, and J. Tsujii. Lexicalized grammar acquisitionn. In *Proc. of the 10th EACL*, 2003. To appear.

[10] T. Hara, Y. Miyao, and J. Tsujii. Clustering for obtaining syntactic classes of words from automatically extracted LTAG grammars. In *Proc. of TAG+6*, pages 227–233, 2002.

[11] XTAG Research Group. A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS-01-03, IRCS, University of Pennsylvania, 2001.