日本語格解析において問題となり得る諸現象の定量的分析

花岡 洋輝 † 松崎 拓也 †
宮尾 祐介 ‡ 辻井 潤一 †§¶
† 東京大学大学院情報理工学系研究科 ‡ 国立情報学研究所コンテンツ科学研究系
§School of Computer Science, University of Manchester ¶National Centre for Text Mining

{hkhana,matuzaki}@is.s.u-tokyo.ac.jp
yusuke@nii.ac.jp tsujii@is.s.u-tokyo.ac.jp

1 はじめに

京都大学テキストコーパス 4.0 (KTC4)[1]，NAIST テキストコーパス (NTC)[2, 3] など，意味情報を付与したコーパスの蓄積が進み，日本語に対する意味解析への関心が高まっている．特に，KTC4 及び NTC は京都大学テキストコーパス [4] において係り受け構造が付与されたものと同一のテキストを対象としていることから，コーパス上で構文構造と意味構造の関係を直接観察することが可能であり，文法主導で構文解析と意味解析を並行的に行う，いわゆる深い構文解析に基づく意味解析をコーパスに基づいて行うためのリソースとしても有用である．本稿では，そのような文法主導の意味解析を行う場合に問題となりうる統語現象について，特に述語項構造と構文構造の関係に着目しながら定量的に分析する．なお本稿では KTC4.0, NTC1.5 を対象として分析を行った．

2 背景

2.1 京都大学テキストコーパス

毎日新聞 95 年度版の記事 38,400 文に対して，各文節の係り先と，各形態素の品詞・活用情報が付与されたコーパスである．例えば，

(1) 国連改革を前提に考えていく

という文については，図 1 のようなアノテーションが付与されている．

2.2 NAIST テキストコーパス

京都大学テキストコーパス中の全記事に対して，述語項構造と照応・共参照が付与されたコーパスである．述語項構造として，述語と事態性名詞について，主格 (ガ格)・対格 (ヲ格)・与格 (ニ格) の格要素が記述され，照応・共参照関係として，ゼロ照応・外界照応も含めた指示詞の照応関係と，名詞間の共参照関係が記述されている．例えば先ほどの文例に対しては，

国連　id=“90”
改革　id=“96”/o=“90”/type=“noun”
考え　o=“96”/o_type=“zero”/type=“pred”

という情報が付与されている．

NAIST テキストコーパスは IPADIC の品詞体系に基づいてアノテーションされているので，本稿でも，IPADIC 品詞体系に基づいて実験・分析を行った．

2.3 Extended Domain of Locality

組合せ範疇文法 (CCG)[5] や木接合文法 (TAG)[6] では，一つの構成素が文脈自由文法よりも大きな制約範囲を持つ (Extended Domain of Locality[7])．この範囲は文法の生成能力に大きな影響を与えるため，構文解析を行う上で重要な概念である．本稿ではその指標の一つとして **spine** を用いる．spine は，構文木上での，ある葉ノードからその葉ノードの最大投射ノードまでのパスとして定義する．これを用いて，構文木上のノード i, j 間の距離 $d(i, j)$ を次のように定義する．

$$d(i,j) = \begin{cases} 0 & (\text{同一 spine}) \\ 1 & (\text{隣接 spine}) \\ \min\limits_k d(i,k) + d(k,j) & (\text{それ以外}) \end{cases}$$

この距離は格関係の複雑さを表す指標となっている．次節で CCG[1] を用いて例を示す．

[1] 本稿では単純化された CCG を用いる．より実用的な日本語 CCG としては戸次の理論 [8] が挙げられる．

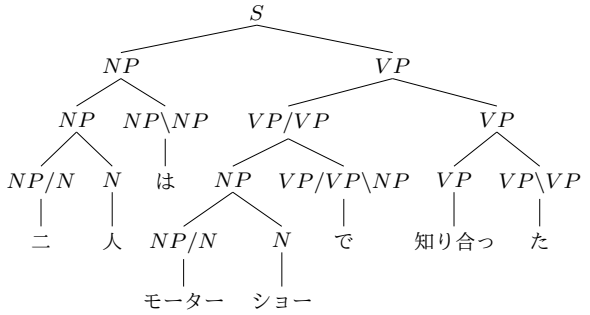
```
* 0 1D
国連/名詞/組織名　改革/名詞/サ変名詞　を/助詞/格助詞
* 1 2D
前提/名詞/普通名詞　に/助詞/格助詞
* 2 -1D
考えて/考える/動詞/母音動詞/タ系連用テ形　いく/接尾辞/動詞性接尾辞/子音動詞カ行促音便形/基本形
```

図 1: 京都大学テキストコーパスアノテーション例

2.3.1 CCG による解析

(2) 二人はモーターショーで知り合った。

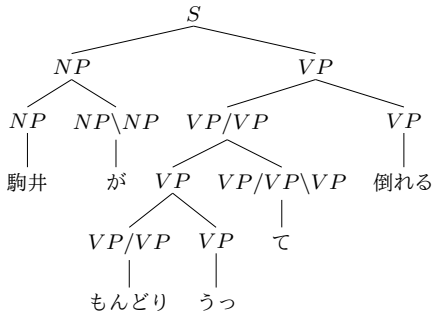
という文を CCG で解析すると (以下，VP は $S\backslash NP$ の略記)，



“知り合っ” とその主格要素である “人” の距離は 1 であり，関数適用規則だけを用いて自明に解析できていることが分かる．次に，

(3) 駒井がもんどりうって倒れる。

という例文を考えると，CCG による単純な解析は，



のようになるが，この解析では “うっ” の主格要素が “駒井” であるという情報を上手く処理できない．この関係を処理するためには，例えば関数交差置換規則のような仕組みが必要である[2]．例のように，一般的には距離が 2 以上の場合には格解析が複雑化すると言える．

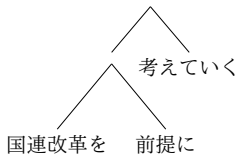
[2]詳しくは戸次の理論を参照のこと．

3 実験

工藤らの研究 [9] において学習データとして使用されている記事 24,283 文を木構造に変換し，一つ一つの格関係に対して，述語 (あるいは事態性名詞) と項の間の木構造上の距離を測定した．

3.1 木構造への変換

小嶋らのアルゴリズム [10] に基づいて，係り受けの構造を，文節を葉ノードとする木構造に変換する．文例 (1) の場合には，

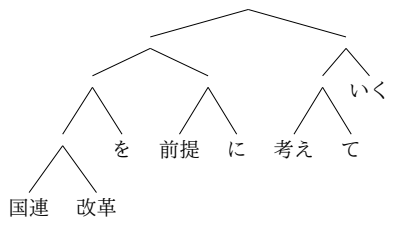


のようになる．

文節内部の木構造については，NAIST テキストコーパスには CaboCha[11] の出力結果に基づく主辞形態素 id が付与されているので，これを手がかりとして，以下のように構築を行う．

1. 文節内部の最初の形態素から主辞形態素までで右下がりの木を作る．
2. 1 で作られた木と残りの形態素から左下がりの木を作る．

文例 (1) の場合は，各文節の主辞形態素はそれぞれ “改革”，“前提”，“考え” であるので，文節内部についても木構造を作ると次のようになる．



3.2 結果

述語についての結果を表 1 に，事態性名詞についての結果を表 2 に示す．これらの表において，“異なる文内” とは，格要素の指定がコーパス中の別の文内にあるものを，“exo?” とは，格要素がコーパス中には出現しないもの，すなわち外界照応を表す[3]．この結果から，

- 異なる文の情報を使わず一文内で解析可能な格関係は，述語で 80%，事態性名詞では 62%程度．
- 同一文内に限定すると，単純な 1 ホップの係り受けで解析可能な格関係は，述語で約 75%，事態性名詞で約 52%．

ということが分かる．

また，同一文節内に述語と項が存在するとき，その間の距離が 0 となるように距離の定義を変更して同一の実験を行った結果を表 3, 4 に示した．この距離を以下 d' と書く．

$$d'(i,j) = \begin{cases} 0 & (\text{同一文節}) \\ 0 & (\text{同一 spine}) \\ 1 & (\text{隣接 spine}) \\ \min\limits_k d'(i,k) + d'(k,j) & (\text{それ以外}) \end{cases}$$

結果を見ると，事態性名詞では距離 0 の格関係の割合が増加している．すなわち，事態性名詞は同一文節内に格関係を持つものが比較的多く，正しく格解析を行うためには複合名詞の内部構造解析が重要になると考えられる．

それに対して，述語では距離 1 の格関係の割合が増加している．詳しく見てみると，距離が 1 に変化したものの内，格関係を記述されている述語が主辞になっていない場合が 8 割程度を占める．すなわち，

(4) 行方が分からなくなった (主辞は “なっ”)

のように述語が複数の動詞からなる場合が多い．日本語では述語に続く機能表現が頻出するため，実用的な文法理論はこれを上手く扱えることが肝要である．

3.3 述語の分析

前節で簡単に分析した距離 $d' = 1$ のものに加えて，距離 $d' = 2$ のものを被覆できれば同一文内の格関係については 9 割以上を被覆することができる．距離 $d' = 2$ の統語現象の約 6 割は，項のコントロール関係に関わる現象である．内訳を細かく見ると，

項 - 述語 1 - 述語 0	1,626
項 - 述語 0 - 述語 1	2,221
述語 0 - 項 - 述語 1	389
述語 1 - 項 - 述語 0	3
述語 0 - 述語 1 - 項	408

ここで “述語 0” は注目している距離 $d' = 2$ の述語を，“述語 1” は同じ項をとる述語を表しており，各行は述語と項がどの順序で並んでいるかに対応する．例えば文例 (3) のように先行述語の格要素が隠れてしまうものは “項-述語 0-述語 1” に属す．他に，

(5) 「司法離れ」は進み、病理的にまでなっている

のように後続述語の格要素が隠れてしまうものは “項-述語 1-述語 0” に，

(6) 連続して地震があった

のように項が先行述語と後続述語の間に入るようなものの多くは “述語 0-項-述語 1” に属す．

これらを CCG で解析することを考えた場合，先行述語の格要素が隠れるものについては関数交差置換規則を用いるなどすれば解析が可能であるが，後続述語の格要素が隠れる場合や格要素が述語の間に入る場合には解析が比較的難しい．

4 おわりに

NAIST テキストコーパス 1.5 に含まれる述語項構造について，関係の複雑さを述語と項の構文木上での距離を用いて数値化し，その頻度を計測した．その結果，事態性名詞に関しては複合名詞の内部構造解析が重要であり，述語に関しては項のコントロール関係を扱うことで同一文内の格関係の多くを被覆できそうであることが分かった．本研究で得られた知見は，文法主導での深い構文解析器を構築していく際の意味解析精度の上限を示すものであり，精度の高い意味解析を行うためにはある程度複雑な文法が必要になることを示唆している．

参考文献

[1] 河原大輔, 黒橋禎夫, 橋田浩一. 「関係」タグ付きコーパスの作成. 言語処理学会第 8 回年次大会発表論文集, pp. 495–498, 2002.

[3] NAIST テキストコーパスにおける exo1，exo2，exog を合わせたもの．

表 1: 述語の格関係頻度

距離	出現頻度	百分率	同一文内
0 (自己参照)	54	0.05%	0.07%
1	61,801	59.63%	74.59%
2	9,964	9.61%	12.03%
3	6,748	6.51%	8.14%
4	2,056	1.98%	2.48%
5 以上	2,230	2.15%	2.69%
異なる文内	9,089	8.77%	
exo?	11,704	11.29%	
total	103,646		

表 2: 事態性名詞の格関係頻度

距離	出現頻度	百分率	同一文内
0 (自己参照)	430	1.03%	1.68%
1	13,461	32.30%	52.49%
2	5,225	12.54%	20.37%
3	2,588	6.21%	10.09%
4	1,547	3.71%	6.03%
5 以上	2,395	5.75%	9.34%
異なる文内	5,673	13.61%	
exo?	10,355	24.85%	
total	41,674		

表 3: 述語の格関係頻度 (文節内距離を 0 とした場合)

距離	出現頻度	百分率	同一文内
0	625	0.60%	0.75%
1	67,961	65.57%	82.03%
2	7,544	7.28%	9.11%
3	3,299	3.18%	3.98%
4	1,675	1.62%	2.02%
5 以上	1,749	1.69%	2.11%

表 4: 事態性名詞の格関係頻度 (文節内距離を 0 とした場合)

距離	出現頻度	百分率	同一文内
0	9,493	22.78%	37.02%
1	6,614	15.87%	25.79%
2	4,184	10.04%	16.31%
3	2,167	5.20%	8.45%
4	1,375	3.30%	5.36%
5 以上	1,813	4.35%	7.07%

[2] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pp. 132–139, Prague, Czech, June 2007.

[3] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 25–50, April 2010.

[4] Sadao Kurohashi and Makoto Nagao. Building a japanese parsed corpus while improving the parsing system. In *Proceedings of the NLPRS*, pp. 719–724, 1998.

[5] Mark Steedman. *The Syntactic Process*. The MIT Press, 2000.

[6] Aravind K. Joshi and Yves Schabes. Tree-adjoining grammars and lexicalized grammars. Technical report, University of Pennsylvania, March 1991.

[7] Aravind K. Joshi. Domains of locality. *Data & Knowledge Engineering*, Vol. 50, pp. 277–289, 2004.

[8] 戸次大介. 日本語文法の形式理論 - 活用体系・統語構造・意味合成. くろしお出版, 2010.

[9] 工藤拓, 松本裕治. 相対的な係りやすさを考慮した日本語係り受け解析モデル. 情報処理学会論文誌, Vol. 46, No. 4, pp. 1082–1092, April 2005.

[10] 小嶋大起, 戸次大介, 宮尾祐介, 辻井潤一. 日本語 CCG の語彙項目獲得. 情報処理学会研究報告, Vol. 2006-NL-176, pp. 75–80, 2006.

[11] Taku Kudo and Yuji Matsumoto. Japanese dependency analyisis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, 2002.