用例の出現頻度を用いた４名詞「の」型名詞句構造解析

佐藤弘幸　　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

日本語文には様々な名詞句が出現し、その構造や意味も多岐にわたる。その中でも代表的な形式のひとつに、いくつかの名詞を助詞「の」によって結合した「の」型名詞句がある。「の」型名詞句は単純かつ基本的なものでありながら、「の」によって作られる名詞同士の関係の多様さゆえに、名詞間の係り受け構造に曖昧性が生じやすい。「の」型名詞句を構成する名詞数が多くなればなるほど、名詞間の関係は複雑になり、構造的曖昧性も増していく。しかし従来の研究では名詞数が３名詞に限定されてるものがほとんどで、４名詞以上の「の」型名詞句について扱ったものはほとんどない。本研究では、「の」型名詞句の用例データベースを用いて、４名詞の「の」型名詞句における構造解析法を提案し、その有効性について検証する。

2　４名詞「の」型名詞句

2.1　「の」型名詞句とその構造

複数の名詞を助詞「の」によって結合した名詞句は一般に「の」型名詞句と呼ばれている。「の」型名詞句は、「の」によって作られる名詞同士の関係の多様さゆえに、係り受け構造にも曖昧性が生じやすく、名詞数が多くなればなるほど係り受け構造は複雑さを増していく。係り受け構造に曖昧性がある「の」型名詞句のうち最も基本的なものは３名詞からなる「N_1の N_2の N_3」という形式のものである。この型の名詞句は、N_1 が N_2 に係る場合と N_2 を飛び越えて N_3 に係る場合があるが、２種類しか存在しないために従来の研究でもかなり高い解析結果が得られている。それに対して「N_1の N_2の N_3の N_4」という形式の４名詞からなる名詞句は係り受け構造が５種類も存在する。

2.2　係り受け構造コード

４名詞からなる「の」型名詞句の５種類の係り受け構造を区別し、扱いやすくするために、係り受け構造コードを定義する。

「N_1の N_2の...の N_k」において

N_iが N_j に係る場合左から i 番目の値は j になる。

N_k はどこにも係ることはないので N_k 自身に係るとみなす。

このコードを使うと４名詞からなる「の」型名詞句５種類は図１のように表せる。

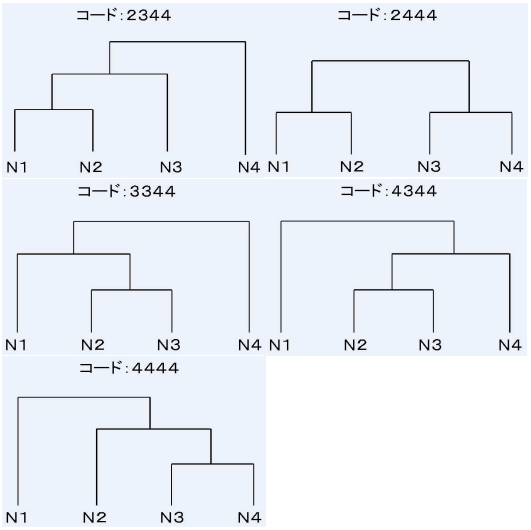

図 1: 係り受け構造

3　接続強度行列を用いた解析

3.1　名詞の分類

名詞は、その働きの違いにより、普通名詞やサ変名詞などといった多くの種類に分けることができる。形

態素解析用辞書においては、名詞は 13 種類に分類されているが、ここでは「の」型名詞句の構造解析に必要な統語的特性を考慮して表 1 のような 7 種類にまとめた。(表 1)。

表 1: 名詞の分類

グループ	品詞	品詞コード
N_n	普通名詞	1100
	形容詞転生名詞	1510
	形容動詞転生名詞	1520
	固有名詞	19**
	代名詞	1b**
N_v	サ変名詞	1210
	連用形名詞等	1220,1230
N_a	状態名詞	1320,1330
N_r	連体詞型名詞	1600
N_d	数詞	17**
	副詞型名詞	1820,1830
N_t	時詞	1810
N_f	形式名詞	1a**

3.2 「の」型名詞句用例データベース

「の」型名詞用例データベースとは 2 名詞からなる「の」型名詞句 N_Aの N_B 約５０万種類を、新聞の記事データから抽出して構築したものである。レコードの内容としては左側表記、左側品詞、右側表記、右側品詞、頻度の５つがある。具体的なレコード形式は表 2 に示す。表 2 の「データ型」は postgreSQL のデータベースで用いられる形式である。品詞コードは１６進数表記をそのまま文字列として登録してある。

表 2: 用例データベースのレコード形式

列名	データ型	例
左側表記	text	大学
左側品詞	text	1100（普通名詞）
右側表記	text	学生
右側品詞	text	1100（普通名詞）
頻度	int	3

3.3 品詞の出現頻度分布

品詞によって助詞「の」の左側に出現しやすいもの、すなわち係り側になりやすいものと助詞「の」の右側に出現しやすいもの、すなわち受け側になりやすいものがあるのではないか、また左右の品詞の組み合わせにも傾向があるのではないかと考えて、「の」型名詞句用例データベース約５０万語の左右の品詞の出現頻度分布を抽出した。結果を表 3 に示す。行が左側品詞、列が右側品詞である。

表 3: 出現頻度分布

	N_n	N_v	N_a	N_r	N_d	N_t	N_f
N_n	273661	166163	1661	5458	12365	9382	14802
N_v	56526	32917	402	960	2006	2290	3399
N_a	2615	1020	18	34	119	38	43
N_r	24114	7264	167	177	393	452	669
N_d	25057	8014	121	291	1275	717	650
N_t	33543	17090	196	825	1175	5031	841
N_f	3555	1587	33	52	59	77	152

3.4 接続強度行列

接続強度行列は「の」型名詞句用例データベースの左右の品詞別の出現頻度分布から求めた７×７の行列である。強度は５から０の６段階である。この接続強度行列は常用対数を用いて算出した。算出例：左側品詞：N_n、右側品詞：N_v
→ 出現頻度：１６６１６３
$log_{10}(166163) = 5.22$ → **強度５**
実際の接続強度行列を表 4 に示す。

表 4: 接続強度行列

	N_n	N_v	N_a	N_r	N_d	N_t	N_f
N_n	5	5	3	3	4	3	4
N_v	4	4	2	2	3	3	3
N_a	3	3	1	1	2	1	1
N_r	4	3	2	2	2	2	2
N_d	4	3	2	2	3	2	2
N_t	4	4	2	2	3	3	2
N_f	3	3	1	1	1	1	2

3.5　接続強度による評価方法

４名詞からなる「の」型名詞句「N_1の N_2の N_3の N_4」において、名詞 N_a はどの名詞に最も係りやすいのか、という観点に基づいて、N_1 から順に係り先を探す方法をとることにした。それぞれの接続強度を接続強度行列から求め比較し、最も強度が大きいところに係るものとする（図 2)。強度が等しいときは名詞同士は距離が遠くなるほど係り受けしにくくなるという性質を考慮して、近いものに係るとした。

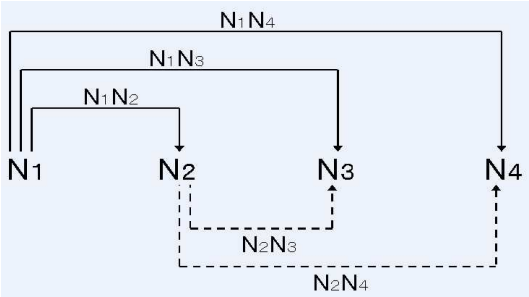

図 2: 解析方法

4　ルールによる補強

用例データベースの出現頻度から求めた接続強度は、あくまで統計情報だけによるものなので、細かい所で必ずしも正しいとは言えない部分がある。そこで各種ルールにより統語的に補強を行う。

4.1　品詞連続ルール

時詞と数詞はその品詞が連続して現われる場合、係り受け関係があることが極めて多い。そこでこの２品詞が連続して現われた場合は必ず係り受け関係を持つものとする。
例：今年度 N_t 年末年始 N_t 海外旅行者 N_n 総数 N_n（係り受けコード ：2344）

4.2　形式名詞ルール

形式名詞はその性質上、直前の名詞句と係り受け関係があることが極めて多い。そこで N_2、N_3 に形式名詞が現れた場合、直前の名詞（N_1、N_2）は必ず形式名詞に係るものとする。
例：細胞 N_n 中 N_f 小器官 N_n 一種 N_d （係り受けコード ：2344）

4.3　時詞ルール

時詞が名詞句の先頭（N_1）に現われる場合、時詞は名詞句全体に係っている場合が多い。そこで N_1 に時詞が現われた場合、N_1 は N_4 に係るものとする。ただし形式名詞ルール等、他のルールによって N_2 に係ると判定されている場合は、その限りではない。
例：九三年度 N_t 六十五歳以上 N_d 高齢者 N_n 転入 N_v（係り受けコード ： 4344）

5　係り受け判定表による解析

5.1　係り受け判定表

用例データベースから得られた出現頻度情報と各品詞の統語的性質を考慮して係り受け判定表を作成した。これにより品詞に関するルールを別扱いすることなく、ひとつの表だけで判定を行うことができる。実際の係り受け判定表を表 5 に示す。

表 5: 係り受け判定表

	N_n N_v N_f	N_a N_r N_{d8}	N_{d7}	N_t
$N_nN_vN_f$	001	000	000	000
$N_aN_rN_{d8}$	001	000	000	000
N_{d7}	001	000	100	000
N_t	011	010	010	100

行が左側品詞、列が右側品詞を表す。品詞連続ルールのために N_{d7}（数詞)、N_{d8}（副詞型名詞）を区別した。この表のフラグの意味は以下のとおり。
０００：係らない
００１：係る
０１０：N_4 に係る（時詞ルール用）
０１１：N_t が N_1 なら N_4 に、N_t が N_2 なら N_3 に係る。（時詞ルール用）
１００：品詞連続ルールによって係る

5.2 係り受け表による解析方法

係り受け表は隣同士の係り受け関係の判定だけに用いる。N_1N_2、N_2N_3 の２つの係り受け関係をこの表で調べることによって、係り受け構造を一意に決定することができる。N_1N_2、N_2N_3 のフラグの組み合わせによる解析パターンは表 6 のようになる。

表 6: 解析パターン

係り受け判定	N_1N_2 フラグ	N_2N_3 フラグ
２３４４	００１	０＊１
	００１	１００
	１００	０＊１
	１００	１００
２４４４	００１	０＊０
	１００	０＊０
３３４４	０００	０＊１
４３４４	０１＊	０＊１
	０＊０	１００
	０１＊	１００
４４４４	０＊０	０＊０
	０１＊	０＊０

6 特定語に対する特別ルール

ある特定の名詞はその品詞の係り受け傾向と異なる傾向を示すものがある。「倍」「前」「後」など、およびそれらを接尾辞としてもつ名詞は、数詞や時詞であるにも係わらず、直前の名詞との係り受け関係があることが多い。そこで、N_2、N_3 にこれらの名詞が現れた場合、直前の名詞（N_1、N_2）は必ずこれらの名詞に係るものとする。

例：北九州 N_n 施設 N_v 十二倍 N_d 規模 N_n （係り受けコード：2344）

また名詞「際」は形式名詞でありながら時詞要素も含むため、時詞と同じルールを適用する。具体的には N_1 もしくは N_2 に「際」が現われた場合、N_4 に係るものとする。

例：入札 N_v 際 N_f 業者 N_n 指名基準 N_n（係り受けコード：2444）

7 解析結果

「N_1の N_2の N_3の N_4」型名詞句２３４例を用いて評価実験を行った。評価実験は出現頻度そのものを接続強度として使用した場合、出現頻度から算出した接続強度を用いた場合、係り受け判定表を使って解析した場合、係り受け判定表に加えて特定語ルールを適用した場合の４種類の方法で行った。結果を表 7 に示す。

表 7: 評価結果

	頻度	強度	判定表	特定語
2344 正解率	61	57	70	72
2444 正解率	31	52	60	68
3344 正解率	25	31	39	53
4344 正解率	0	0	37	35
4444 正解率	12	25	37	42
合計正解率	40	47	55	61

実験の結果、最終的な数値として６１％という結果が得られた。この数値はまだ不十分な結果である。しかし３名詞の場合の係り受け構造が２種類しかないのに対して、４名詞では５種類の係り受け構造が存在する。その点を考慮すると本手法は、４名詞の解析にある程度効果を示していると言える。

8 おわりに

本稿では、用例データベースを用いて品詞ごとの係り受け関係を分析し、品詞による４名詞「の」型名詞句の構造解析法を提案し、その有効性を示した。しかし一方で４名詞以上の「の」型名詞句は３名詞の場合とは比較にならないほど構造の複雑さが増しているので、品詞だけによる解析では限界があることも確かである。今後は意味情報など品詞以外の要素にも目を向けさらなる解析率の向上を目指したい。

参考文献

[1] 益田裕也、宮崎正弘：名詞間の接続強度を用いた「の」型名詞句構造解析
言語処理学会第９回年次大会発表論文集、C3-1(2003-3)

[2] 武本裕、宮崎正弘：名詞間の接続強度と用例の係り受け情報を用いた「の」型名詞句構造解析
言語処理学会第１０回年次大会発表論文集、C5-1(2004-3)