第 5 章

統計的手法を用いた日本語形態素解析の曖昧さ解消方式

5.1 はじめに

日本語の形態素解析という処理の問題について考えてみると、一般に、汎用性の高いシステムほど解析可能な文章が多くなる。しかしながら、より多くの日本語文を解析するためには、様々な解析ルールを用意する必要があり、その解析ルールが可能性のある全ての解析結果を出力すると、解析結果の数が膨大になるという欠点をもっている。つまり、より汎用性の高いシステムをめざすと、そこに必然的に曖昧さの問題が生じてくる。本章では、曖昧さを除くための一手法として、大量の言語データを収集して分析し、言語自体に対する知見を実証的に得ることを試みたので、それについて述べる。

これは構文解析の段階の曖昧さの爆発を防止するため、その前の形態素解析の段階での曖昧さを予め減らそうとするもので、新聞記事から曖昧な語句を前後の文脈と共に抽出し、人手で分類・整理し、曖昧さを解消する方式について検討した。

5.2 形態素解析における文節の曖昧さを含む語句の分類

日本語の形態素解析で問題とされる曖昧さには各種のものがあるが、その中の代表的なものに格助詞あるいは助動詞「で」の曖昧さがある。例えば、

- 本<u>で</u>殴った。

という場合には「で」は格助詞である。しかし、

- これは本<u>で</u>、あれはペンだ。

という場合の「で」は助動詞となる。このような「で」を、「で」の直前に現れる名詞の意味カテゴリ (SF:Semantic Feature) で見分ける方法もあるが、

a) 「道で（会う）」

b) 「収賄の容疑で（あげられる）」

c) 「ストライキで（足を奪われる）」

等々、直前に現れる名詞の意味カテゴリも場所、時、人称、状況、理由、道具、など多岐にわたり、その対応関係は複雑である。

本章では、

(1) 浅い知識のみを用いてできる限り形態素解析の曖昧さの数を減らし、

(2) 残された可能な解析結果についても優先度を付与することによって、より尤もらしい解析結果が選択できるようにする

ことを目指している [40]。そのため、格文法でなく結合価文法を導入し、日本語文解析の尤もらしさの目安としている。本章では、(1) の実現のため、大量のデータをもとに形態素解析プログラムに制限を加えることを試みたのでその方法と結果について述べる。

前述したような「で」は本来形態素解析の段階で上記 a)、b)、c) の区別がつくことが望ましいが、現在の計算機による自然言語処理ではなかなか容易でない。これは形態素解析の段階では字面だけの処理が中心で、上記の区別をつけるためには意味処理を先行させる必要があるためである。

「で」は前述のように、格助詞と助動詞の曖昧さをもち、「に対（する）」「に関（する）」は、まとまって1つの格助詞相当の連語（以後、これを「格助詞相当連語」と呼ぶ）となる場合と「対（する）」「関（する）」が本動詞として用いられる場合の2通りの解釈ができる。本章で述べる統計的な手法による日本語形態素解析の曖昧さ解消の目的は、人手でこれら「で」、「に対する」および「に関する」を分類し、形態素解析の時点で品詞を決定し、さらにその用法上の特徴を見出すことである。

5.3 提案する曖昧さ解消方式

まず、第1段階として前出の「で」と、「に対（する）」「に関（する）」について調査を行った。

新聞記事1ヵ月分について調査を行った結果、「で」はごく一部の例外を除く、以下の場合が格助詞、それ以外の場合は助動詞と解釈してよいことが判明した。

「で」が格助詞となる場合

① 「で」＋動詞

② 「で」＋助詞

ここで、例外となるのは、①の場合、

＜例6．1＞「大変な努力が必要で疲れてしまう。」

②の場合、

＜例6．2＞「始めからあったわけではない」

①の例外は、助動詞「だ」の連用形「で」であり、本来「、(読点)」等で中止されるべき文章が、「、(読点)」を伴わずに続けて書かれた場合である。また、②の例外は、「～ではない」など、あるいは、「～である」または「～である」の変形が後ろにくる場合である。

表 5.1: 「に対（する)」の調査結果

使用例	格助詞相当連語	本動詞	計
に対し	545	0	545
に対して	239	0	239
に対しては	65	0	65
に対しても	35	0	35
計	884	0	884

表 5.2: 「に関（する)」の調査結果

使用例	格助詞相当連語	本動詞	計
に関し	18	0	18
に関する	196	0	196
に関して	56	0	56
計	270	0	270

一方、「に対（する)」は、変形（「に対して」、「に対しても」等）を含め、新聞1か月分に出現した884例のすべてが格助詞相当連語であり、「対（する)」が本動詞として用いられる例は1例もなかった（表 5.1 参照)。また、「に関（する)」についても270例のすべてが格助詞相当連

語として使用されており、本動詞としての使用例は1例も発見されなかった（表 5.2参照）。そこで、今後は「対（する）」と「関（する）」には動詞の解釈をしないこととする。

以上述べたような方法で、「で」と「に対（する）」「に関（する）」という3種類の語句の形態素解析における曖昧さを解消することを試みた。他の語句については、添付資料 A.1.2参照のこと。

5.4 形態素解析における品詞多義の問題

次に第2段階として、様々な役割を担う助詞「と」について分類および考察した(参考資料 A.1.3)。

「と」は、

a) 再続投する考えはあるかと の質問に答える。 ［格助詞］

b) 夜になると 出てくる。 ［接続助詞］

c) いきいきと ふるまう。 ［慣用表現］

のように、大きく分けて3種類の「と」があると考えられる。一見して格助詞、接続助詞、慣用表現と見分けのつくものもあるが、そうでない「と」も数多い。

「と」は大別すると前述のように3種類であるが、ここではその後の一般的な日本語処理のことを考え、さらに細分類して次に示す9種類に分類することとした。

(1) 格助詞

(a) 引用：いわゆる引用のほか、疑似引用等も含む。

(b) 並列：英語にすると"and" の意味のもの。

(c) 随伴：英語にすると"with"の意味のもの。

(d) 結果：「と」を受ける動詞の主語が、「と」の前にくる名詞と一致するもの。

(e) 格助詞相当連語「～として」、連体修飾表示「～としての」の「と」。

(f) 補助動詞「する」の直前の「と」。

(2) 接続助詞

(3) その他

表 5.3: 「と」の分類

種類		頻度	割合
(1)	(a)	3353	43.1
	(b)	1061	13.7
	(c)	919	11.8
	(d)	619	8.0
	(e)	540	6.9
	(f)	180	2.3
(2)		814	10.5
(3)		288	3.7
	合計	7774	100.0

(a) 『状況化』の「と」

(b) 慣用的表現に含まれる「と」: 分類不可能ではないが、まとめて慣用的に扱えるもの。

上記の分類基準に基づき、実際の新聞記事データにあたり、各々の出現頻度について表 5.3にまとめた。

今回の調査は朝日新聞の朝・夕刊を合わせて1ヵ月分の文章を対象とし、形態素解析を行った結果、格助詞、接続助詞についてはその品詞番号を伴った「と」を抽出して分類した。一方、格助詞、接続助詞以外の「と」については、すべての「と」をKWICにより出力して、人手により例外を取り除き、その上で今回の分類上、どのカテゴリに含まれるかを調べた。

この表を見ても明らかなように、引用(1). (a)、並列(1). (b)、随伴(1). (c)といった、"格助詞らしい"「と」の使用率が非常に高く、これら3つで全体の約7割を占めている。特に引用(1). (a)の割合が高いが、これは今回調査した対象となる文書が新聞であることと大きく係わっていると考えられる。

同時に、これらの「と」を受ける動詞についても調査したところ、必ず「相手」を必要とする動詞(ここでは随伴性の動詞と呼ぶ:「争う」「結婚する」「違う」など)が「と」の後ろに続いた場合、その「と」は随伴の働きをする確率が非常に高いことがわかった。

反面、随伴の「と」が出現した場合の動詞について見てみると、この場合はほとんど動詞を選ばず、本来相手の必要でない動作に対しても臨時に相手を伴う、つまり臨時に随伴性を帯びることがあることもわかった(例6.3参照)。

<例6.3> 昨日は1日あの人と飲んだ。

この場合、飲みに行くのに必ずしも相手は必要ないが、「飲むという行為」を共にするという意味で臨時に随伴性を帯びているのである。

ここで、引用の「と」において『引用性』というものを考えてみると、「いう」「考える」「述べる」など、もともと「句＋と」を受けるのが自然な動詞（例6．4参照）と、不自然な動詞（例6．5参照）が存在する。不自然とみられる動詞が「句＋と」を受けている場合は臨時に引用性を帯びていると考えられるが、これらを機械処理しようとする場合、臨時に引用性を帯びる、いわば疑似引用ともいうべき動詞をどのように扱うか、問題が残る。

＜例6．4＞　5人にひとりが高齢者<u>と</u>いう6町なのだ。

＜例6．5＞　桜の花びらが雪<u>と</u>降る。

しかし、新聞記事1ヵ月分に出現した、引用あるいは随伴の「と」を受ける動詞、のべ8676個、異なり語数633個の動詞の中で、出現数4336回、しかも随伴の意味の「と」を1回も受けなかった動詞「いう」や、出現数597回で同じく随伴の意味の「と」が1回も現れなかった動詞「思う」など、非常にはっきりした性質を示した動詞については、「と」の役割を限定して構文解析をすすめてもよいと考えられる。

(1). (d)の「結果」の意味の「と」は、一般に「転化の結果」といわれるものである。今回ここでは、「『と』を受ける動詞の主語が、『と』の直前にくる名詞と一致するもの」に限定してみた。この条件により、結果の「と」を受けうる動詞は「する」「なる」「映る」「化す」「みえる」の5種類に絞ることができた。

資料A.1.3で1D'としたのは「状態の『と』」ともいえるものである。「山と積み込む」というのは「積み込んだ結果、山となる」の意であって、「結果の『と』」の特殊なものと考えられるが、実際には調査した2万例中、資料にも示した1例のみしか出現しなかった。

その他、特殊な形で出現する「と」、具体的には「と」の直後に「。(句点)」を伴う表現について調査したので、その結果の一部を表5.4に挙げる。

以上みてきたように、この節では助詞の「と」に注目してその分類を試みた。この結果非常に明確な性質をもつ「と」については形態素解析の段階での曖昧さを減らすことが可能であることが判明した。同時に今回の調査によって、いくつかの動作については「と」を受ける場合の意味が明らかな動詞が存在することもわかったので、構文解析の際の解釈の仕方が狭められると考えられる(参考資料A.1.4)。

表 5.4: 「と。」の分類

表現	全数	引用	接続助詞	その他
」と。	２３	２３	0	0
、と。	１４	１４	0	0
ー と。	1	1	0	0
ーー と。	1	1	0	0
…… と。	1	1	0	0
ないと。	8	8	0	0
だと。	3	3	0	0
命令形＋と。	1	1	0	0
その他＋と。	１５	8	2	5
合計	６７	６０	2	5

5.5 統計的手法を用いた日本語形態素解析の曖昧さ解消方式の評価

前節でみてきた形態素解析の曖昧さを減らすための方策を実現し、データとして朝日新聞天声人語６日文の記事について構文解析ルールを適用し、実際にどれくらい曖昧さを減らすことができるかを実験してみた結果について述べる。

対象としたのは資料 **A.2** に添付した朝日新聞の天声人語で、文の数は全体で１５６文である。資料のなかで途中に見られる空白行およびタイトル行は除いてある。文の長さは平均で３４．６文字、最も短い文で７文字、最も長い文で１０４文字あった。文の長さの中央値は３１文字、最頻値は２８文字である。曖昧さ解消のため、具体的には

(1) 形態素のレベルで、各語に付与されている品詞を１つに絞る

(2) 名詞の連続は１語の名詞とし、１番後ろの名詞の意味カテゴリを継承する

(3) 動詞の連続は１語の動詞とし、１番後ろの動詞の意味カテゴリを継承する

(4) 文節内の曖昧さは単語数最小の解釈を選択する

(5) 以下にあげる連語の品詞を固定する

表 5.5: 曖昧さ解消方式適用前後の構文解析の成功率の変化

	文数 S	MA 成功 Ms	MA[1]成功率 Ms/S(%)	MTE Sm	SA[2]成功 Ss	解析木の平均	MA 成功率 Ss/S(%)
適用前	156	130	83.3	32	34	10.3	21.79
適用後	156	130	83.3	11	53	13.8	33.97

にあたって、において、における、に関する、に対し、に対して、に対する
にとって、によって、により、による、によると、によれば

のような処理を行った。

実験は、まず上記曖昧さ解消方式を用いずに従来の方法で形態素解析、構文解析を行って解析された文の数、解析された文の構文解析木の数を求め、次に上記曖昧さ解消のための処理を行った後に、形態素解析、構文解析の結果を調べた。

この実験によって、表5.5に示すような結果が得られた。

ここで、ＭＴＥというのは Machine Time Extended といって、ＣＰＵ時間で１分をかけて解析が終了しないほど曖昧さが大きいということであり、この他に解析木を保存するためのメモリが不足するエラーもあるが、今回の実験ではメモリ不足によるエラーは発生しなかった。

今回の実験では、当初ＭＴＥのために１５６文中３２文が構文解析不能であった。また、主として長い文では形態素解析の段階で既に解析に失敗しているものもあるが、品詞の特定や連語の処理によって構文解析の解析率を上げるという目的は達成できた。

この表で、解析木の平均といっているのは、最終的に構文解析された解析木の数を解析できた文の数で割ったものである。解析木の平均は曖昧さ解消方式の適用前と比べて適用後に数が増えている。これは一見、曖昧さが増えているように見えるが、実際は、曖昧さ解消方式を適用しなかったときには解析できなかったような長い文章が、適用後は構文解析可能になったこと、しかもそのような文はその長さゆえに解析の曖昧さが多くあるため結果として解析木の数が増えることになったためである。

このように上で述べたような曖昧さ解消のための方策をとることにより、構文解析可能率は１０％以上向上したことが分かり、この方法は形態素解析のみならず構文解析の曖昧さ解消にも有

効であることが判明した。

5.6 形容詞の多義を減らすために提案する曖昧さ解消方式

ここまでは実際にシステムで評価を行うまでにいたった各種曖昧さ解消方式について述べてきたが、ここではまだシステムで評価を行うまでには至らないが、構文解析処理における曖昧さ解消のために行った作業について述べる。

(1) 程度副詞になる形容詞の種類と用法調査「すごく」、「ひどく」、「えらく」、「著しく」、「恐ろしく」等の程度副詞としての用法のある形容詞を抽出した。朝日新聞4か月分を対象としたデータから程度副詞としての機能をもつ形容詞を拾いだし、実際にその係り先としてどのような品詞があったかをリストアップする。そして、同じ形容詞の連用形でも程度副詞としてではなく、本来の形容詞として用いられている場合、それはどこに差異があるのか考察した。

(a) 明らかに程度副詞としての職能をもつもの (表 5.6)

(b) おそらく程度副詞としての職能をもつと思われるもの (表 5.7)

表 5.6: 程度副詞として用いられた形容詞一覧-1

形容詞	程度副詞として用いられた例と係り先					形容詞として用いられた例	合計
	形容詞	形容動詞	動詞	名詞	小計		
激しく	1	0	107	11	119	108	203
著しく	10	5	48	2	65	5	70
ひどく	5	4	20	0	29	32	61
すごく	9	1	7	2	19	2	21
ものすごく	6	1	7	0	14	1	15
華々しく	0	0	4	0	4	0	4
すさまじく	0	0	1	0	1	6	7
計り知れなく	1	0	0	0	1	0	1
限り無く	0	0	1	0	1	0	1
合計	32	11	195	15	253	154	383

[1]Morphological Analysis (形態素解析)
[2]Syntactic Analysis (構文解析)

表 5.7: 程度副詞として用いられた形容詞一覧-2

形容詞	程度副詞として用いられた例と係り先					形容詞として用いられた例	合計
	形容詞	形容動詞	動詞	名詞	小計		
恐ろしく*	0	0	0	0	0	3	3
すばらしく	0	0	0	0	0	2	2
目覚ましく	0	0	0	0	0	2	2
甚しく	0	0	0	0	0	1	1
合計	0	0	0	0	0	8	8

*)「空恐ろしく」の場合は必ず「空恐ろしくなる」の形で出現するため程度副詞としての職能はもたないと考える。

(2) 程度副詞になりうる形容詞が程度副詞にならないのは大きく分けて以下の5つの場合である。

(a) 直後に動詞「なる」がくるとき（93例）
例:混雑はいっそうひどくなり、‥

(b) 直後に読点がくるとき　　（54例）
例:個人主義がひどく、思想作風が悪く、‥

(c) 直後に接続助詞「て」がくるとき（5例）
例:もうけたくても競争が激しくて難しい。

(d) 直後に動詞「する」がくるとき　（2例）
例:資金の流れをさらに激しくする。

(e) 直後に（係助詞＋）形容詞「ない」がくるとき（2例）
例:かつてのように激しくはない。

また、程度副詞になりうる形容詞が程度副詞になっていない例は162例あったが、そのうちの156例が上記の5つの場合に含まれていた。例外の残り6例を以下に示す。

例外① ～年、太平洋戦争はいよいよ激しく敗戦の色はこくなった。
例外② カーペットはむしろ、動きが激しく騒音の出やすい営業などの部門に敷くべきだ。
例外③ 頭痛がひどくぐっすり眠れなかった。
例外④ 市内の銃撃戦は日ましに激しく食料の買い出しもできない。
例外⑤ 利副が薄いうえ、競争も激しく「新国鉄が乗り出してくるのは大変な脅威」（交通公社）と緊張している。

例外⑥ このところお互いに健忘症がひどく２人３脚、口げんかもレクリエーションとなり、…

これらは直後に読点を付したほうが自然と考えられる。

＜例外3'. ＞頭痛がひどく、ぐっすり眠れなかった。

ゆえにこれらは、本来続くべき読点が脱落したものと判断できるが、１６２例中の６例ということで、例外とみることにする。

(3) 形容詞の連用形に関するまとめ

形容詞の連用形が程度副詞として働くための条件は以下に示す図 5.1のとおりである。

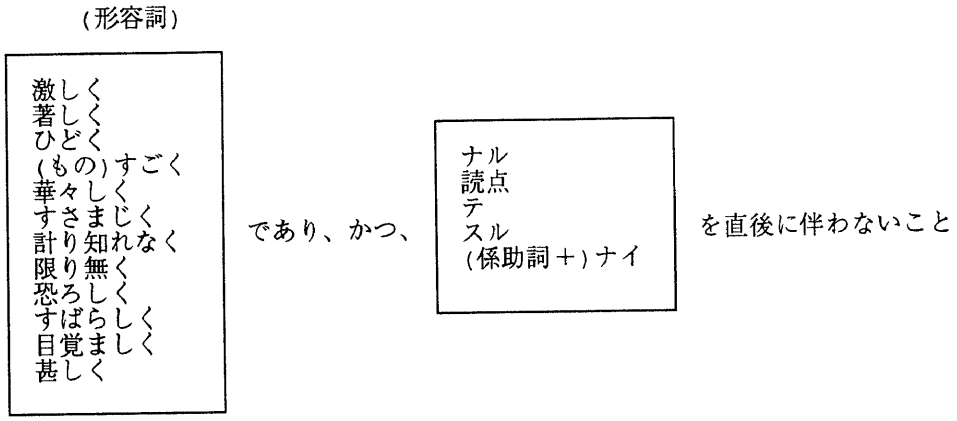

図 5.1: 形容詞が程度副詞として働く場合の条件

5.7 まとめ

以上みてきたように、既に調査を行って検討し、形態素解析ルールとして組み込めるものについてはルール化したが、修飾関係、品詞等、構文解析に係わる曖昧さの解消に役立つと考えられる問題点にはこれ以外にもあるので、重要と考えられるそのうちのいくつかをここであげておく。

(1) 「と」の受けるもの

引用、随伴、並列、接続助詞と、「と」には大きく分けて４つの用法があり、これらを見分けることが構文解析の曖昧さ解消に非常に役立つ。既に一部に関しては調査・検討したが、現在の段階ではルール化するには問題があり、分類基準を明確にするためにはさらに調査・検討が必要である。

(2) 形容動詞語幹と名詞の判別添付資料Aにあげたように、既に調査は行い、辞書登録も済ませたので、今後はこの辞書を用いて実際にどれくらい曖昧さの解消に効果があるかを調べる。

(3) 程度副詞になる形容詞の辞書登録

抽出した「すごく」、「ひどく」、「えらく」、「著しく」、「恐ろしく」等の程度副詞としての用法のある形容詞について辞書登録を行い、その効果を調べる。

(4) 「の」がついて必ず述語になる語の調査

例えば「最大の値」のように、もともとは活用しない語が述語的な職能をもつものと、形容動詞の語幹に「の」がつくものとがある。今回は形容動詞で「の」がつくものについて調査し、辞書登録を済ませた。名詞に関しては未調査である。

(5) 「ハガ文」の調査

いわゆる「象は鼻が長い」という文を典型とする「～ハ～ガ～」という型の文を、機械翻訳しやすい形、例えば「象の鼻は長い」のような文に変形することができないかを調査した。「ハガ文」には非常に多くのタイプがあり、現在までに一部の調査を行ったにすぎないが、今後も調査を続ける予定である。

(6) 動詞連用形＋「に」＋動詞

例えば「釣りに行く」のような形で出現する動詞にはどのようなものがあるのかを調査した。辞書登録を残すのみである。

(7) 感動詞、接続詞を分ける

従来、感動詞、接続詞は辞書の中で同じ接続コードをもっていた。これは形態素解析上は問題にならないが、構文解析の段階では2種類の品詞の係り先が異なることから、曖昧さが増大する。そこでこれらを分けるよう、接続コードを別にすることにした。分類は既に終り、あとは辞書登録を行う。

その他にも、今後調査を要する項目としては、

(1) 「AをBに」という構文で、AがBの格を埋めているケース

例えば、「鐘を合図に行進が始まる」では「鐘が合図になる」ことによって行進が始まったということで、「鐘を」が「合図する」の格を埋めている。

(2) 助動詞及び形容動詞の連用形「で」の分類

- 形容動詞か名詞かの区別

- 形容動詞ならば並列か連用かの区別

(3) 格助詞「の」の取り扱い

上記作業項目 (4) の発展として、「不振の選手」「最大の値」「逆の考え」など、「な」に接続しなくても形容動詞と考えられるものをどのように扱うかについて検討が必要である。

(4) 「と」:「～となる」の「と」

- 「重要となる」の「と」は「重要と言う」の「と」と同じかどうか
- 「重要とする」の「と」は「重要と言う」の「と」と同じかどうか
- 「と」の前に付きうる品詞とその種類の調査
- 「と」の後に付きうる品詞とその種類の調査

(5) 「に」: 格助詞と「～となる」の「と」

- ニ1格とニ2格の推定の基準

などが考えられる。

今後は、上で述べたような修飾関係、品詞等、構文解析に係わる曖昧さを解消するために残されている数多くの問題点を解決するべくさらに調査を行うつもりである。