自然言語処理 77－1
（1990．5．18）

形態素解析システムにおける新聞記事の調査とシステムの評価

関根 聡[1)]　菅野 祐司[2)]　長尾 健司[2)]

1)日本電子化辞書研究所　2)松下電器産業株式会社

日本語処理基本システムの形態素解析部とその辞書を開発し、その評価を行ったので報告する。　形態素解析は基本的に文節数最小法を用い品詞等によってコストをつけ、コストの合計値の少ないもの程、確からしい文とする方法を取っている。本形態素解析は、コスト幅を最小のものから任意幅で任意個の解をグラフの形で出力する。辞書は九州大学の辞書を基本に、固有名詞、カタカナ語等を加え１３万見出しの辞書を開発した。　評価は、新聞記事１カ月分（約７万文）に対して行い、解グラフ生成率、正解率、文長と解析時間の関係、形態素出現頻度、解析失敗の原因、誤答の原因の解析等の評価を行った。

The Linguistic Survey for News papers by Morphological Analysis System and the Evaluation of the System

Satoshi SEKINE[1)]　Yuuji KANNO[2)]　Kenji NAGAO[2)]

1) Japan Electronic Dictoinary Reserach Institute, LTD.
2) Matsushita Electric Industrial CO., LTD.

1) 1-4-28 Mita, Minato-ku, Tokyo 108 JAPAN.
2) 3-10-1 Higashimita, Tama-ku, Kawasaki 214 JAPAN.

We report the linguistic survey for news papers by morphological analysis system and the evaluation of the system. This system is based on the Minimum Phrase Group Method, and introduce the cost which is determined by each categories. We can get graph of phrase as result under any range of costs. The dictionary consists of 130,000 words. It has deveroped mainly based on the dictionary of Univ. KYUUSYUU and we added some words.

The evaluaton was held for news papers for a month (about 70,000 sentense). As a result, we got the rate of generation of graph, the rate of success, the relation between the length of sentense and analysing time, the rate of the number of morpheme and the reason of fail.

1) 本内容は筆者が松下電器株式会社在職中に行った研究を発表したものである
1) This research was made while the auther was belong to Matsushita E.I.Co.Ltd.

1．はじめに

現在、機械翻訳や対話システム、文章処理等の自然言語処理応用システムには、その基本部分の中に共通に利用できる部分が少なくないということが言われている。言い換えると、そのような基本的なシステムをあたかもツールのように予め用意しておけば応用システム開発の労力は大幅に軽減されるということが言える。

そこで我々は、特に日本語処理を対象に基本的なツールとして辞書、形態素解析、構文解析のシステムの開発を行ってきた。[1)2)]我々はこれらを日本語処理基本システムと呼ぶ。このシステムのうち、形態素解析システム、辞書についての開発を行った。本稿では、このうち形態素解析システムとそこに使われる辞書について報告し、形態素解析システムの評価を行ったのでそれについて述べる。

2．日本語処理基本システムの全体構成

まず日本語処理基本システムの全体構成について述べる。この基本システムは、日本語のべた書きの文を入力として形態素解析、構文解析を行い、文の形態素情報、構文情報を出力する。またこのために必要な辞書も基本システムに含む。

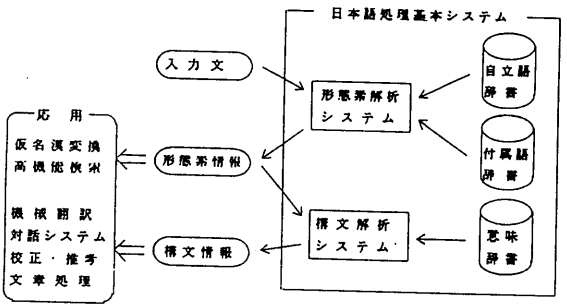

図１　日本語処理基本システムの全体構成

また、開発は以下の基本思想に基づいて行った。

(1)　高精度

解釈の曖昧性の問題に対処するため、可能性のある全ての解釈に対して文のグローバルな情報を用いて尤らしさの判定を行うようにする。

(2)　高速

(1)を実現することは、とかく処理の高速性を損なうことにつながりがちであるが、解析の任意の段階で、複数ある解釈に含まれる共通の要素に対しては、処理を共通にすることによりこれに対処する。

(3)　コンパクト

共通する要素に対する処理の共通化は、おのずとデータ量の削減につながるがこれに加えて各種データを表現する効率的なシステムを構築することにより、所要記憶量の低減をはかる。

3．形態素解析システム

日本語処理基本システムの形態素解析システムについて詳しく説明する。この形態素解析システムは、入力としてべた書き文を入力し、出力は日本語の曖昧性を含んだ形態素グラフを出力する。

このグラフは基本システムで述べたように、コンパクト性を意識し、共通部分は共通なグラフとして、曖昧性のある部分だけを分割した以下のような形になっている。

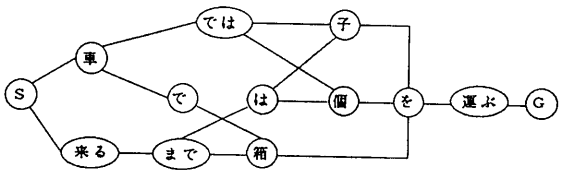

図２　「くるまではこをはこぶ」の形態素解析結果

また、解析のたびにすべての曖昧性を含んだ出力を出すと、全く必要のないものまで含まれてしまう可能性があるので、指定したいくつかの尤らしい解析結果だけを求め出力するようになっている。このグラフ形式の出力は次の構文解析システムにそのまま渡すことが出来る。構文解析では、この曖昧性を含んだグラフを効率よく処理するようになっている。[4)]

3－1　システム概要

本システムの概要を示す

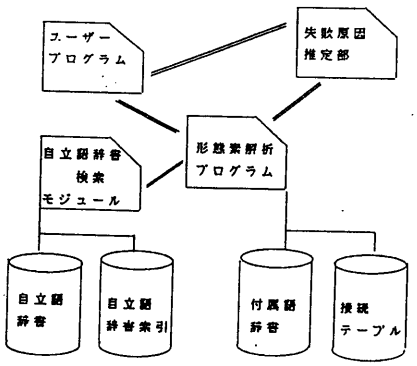

図3　形態素解析システムの概要

本システムは、上記のように4つのプログラムと4つの辞書から成る。このうち中心となる形態素解析プログラムと失敗原因推定部を除いた部分は、フォーマットや関数仕様さえ合わせればこのシステムのもの以外でも使用できる。したがって、少しの変更だけで多くの目的に対応させることができ、ポータビリティーが高いということが言える。また、形態素解析のプログラムのインターフェースは6つの関数から成っておりユーザーがそれらの関数を用いて自由にプログラムして、自分の目的に合ったシステムを作ることができる。

3－2　解析アルゴリズム

一般に、漢字かな混じりの入力文字列に対する辞書検索の結果は複数あり、通常の形態素解析で行うような接続ルールを用いた絞り込みだけでは、これを一通りに決めることはできない。したがって、形態素解析は入力文字列と日本語辞書及び接続ルールによって決まる探索グラフの中で経験的な知識等に照らして、もっともらしい幾つかの解析に対応するサブグラフを探索するという問題として捉えることができる。

そこで本形態素解析システムでは発見的なグラフ探索の手法A*を基本とし、もっともらしい解パスよりなるサブグラフを取り出せるよう拡張しこれをベースにしたアルゴリズムを用いる。[3)]

A*は、与えられた探索グラフからコストが最小である任意のパスをただ一つ探索して停止するものである。[5)6)]我々はこのA*アルゴリズムをもっともらしいパスよりなるサブグラフを取り出せるように拡張した。具体的には探索コストが（最小コスト＋α）以下の範囲内でn番目までの解パスを出力できるように拡張した。またこの拡張による時間計算量は、O（n^2）であり拡張による時間計算量の悪化はほとんど認められないことが証明されている。

次にコストの付け方であるが、上記のアルゴリズムより明かなように、コストは小さい程尤らしさが高くなっている。また、現在形態素解析で主流となっている文節数最小法をこのアルゴリズムにそのまま適用することを考えると自立語のコストを1、付属語のコストを0とすればよい。しかし、実際にインプリメントした場合、コストをそのように単純に付けるのではなく、あるヒューリスティックに基づいて付けた方が解析率がよりよくなることがわかっている。そこで我々は、文節数最小法及びヒューリスティックに基き以下のようにコストを付けた。

・自立語	
・用言語幹	90
・用言語幹を除く自立語	100
・付属語	
・一部の格助詞、並立助詞、文節助詞、一部の副助詞、係助詞、接続助詞	0
・並列助詞、終助詞	9
・接辞	60
・その他の付属語	10

図4　コストの付け方

つまり、形態素解析は、まずべた書きの文字列から辞書検索により単語の切り出しを行う。そしてそのうち接続ルールで接続可能な単語列を抽出しグラフで表現する。そのグラフの中から各単語についたコストの和の小さなものをグラフ探索により求める。このように、べた書き文字列からヒューリスティックなコスト計算の結果、文としてもっともらしい単語列を求めることが形態素解析であると言える。

3－3　自立語辞書検索

自立語辞書の探索は、以下の特徴を持っている。

（1）　かな、混ぜ書き、正書の全ての表記での効率

的検索

(2) 形態素解析の際の検索され方を考慮した効率化

(3) 辞書本体や索引情報の容量のコンパクト化

の特徴を持っている。辞書本体はその容量の大きさを考慮し、２次記憶上に小さなブロック単位で格納され、検索時には検索ブロック限定Ｔｒｉｅを用いて高速な検索ができるようになっている。

３－４　自立語辞書

我々のシステムに含まれている辞書は、約１３万見出しでそのうち基本語部分は、九州大学基本語辞書８万見出しを使用している。また固有名詞として、松下標準固有名詞辞書を利用している。この固有名詞辞書は約２万見出しの姓名、約２万見出しの名前、約１万見出しの地名（国内地名がほとんどで海外の地名は１００以下）約１５００見出しの企業、団体名（企業は東証１、２部上場すべて）が含まれている。このうち、本形態素解析システムの辞書検索に必要な読みの対応情報を自動的に付加した約４万５千見出しを自立語辞書に使用している。また、以上の辞書構成では、カタカナ語、アルファベット語に弱いという欠点があるためそれらについては、新聞記事からの抽出を行い高頻度のものに品詞情報、意味情報を付けて自立語辞書に加えた。その数は、約８０００見出しである。

３－５　付属語辞書、接続表

付属語辞書は、約１３５０見出しから成る。主に活用語尾や助詞から成るが、特徴として接頭辞８４見出し、接尾辞３２１見出し、助数詞９４見出しが含まれており、接辞等が充実している。また、接続表は、二方向接続属性を採用しており、２５１×４６０のマトリックスになっている。このマトリックス中の接続可否の判断はすべて人手によって行った。

４．評価

この形態素解析システムの評価を行った。この評価の目的は、この形態素解析システムの特徴や欠点を見い出すことにある。しかし形態素解析の評価の項目や方法はそれぞれのシステムの特徴や、入力、出力方法の違いにより一意に決まるものではない。そこで今回は、このシステムの特徴等を考慮し、以下のような評価項目と調査項目を設定した。

1) 解グラフ生成率
2) 正解率
3) 文長と解析時間の関係
4) 形態素の出現頻度
5) 解析失敗の原因
6) 誤答の原因の解析

ここで解グラフ生成率とは、すべての解析対象文中このシステムで解のグラフを生成できた文の割合をいい、解析失敗とはグラフの生成ができなかったことをいう。また、正解率は生成されたグラフの中に正解（構文的にも意味的にも人間が見て正しい解釈であると思われる解析結果を生成できたもの）が存在する文のすべての解析対象文の数に対する割合をいう。誤答とは、解析を生成にした文のうち正解が含まれなかったものをいう。形態素の出現頻度とは、解析に成功したすべての文に表れる形態素の出現した回数の頻度を求めたものである。このシステムでは解はグラフとして出力されるため、頻度の求め方は、グラフから解釈可能な解に数に対する出現回数の割合として求める。例えば、以下のようなグラフが解として求められた時は、『辞書』『の』の出現頻度は１『機能』『昨日』の出現頻度は０．５とする。

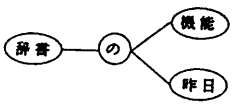

図５　出現頻度の数え方

今回の評価では、解のコスト幅として１０ポイントとした。これは、出力として構文解析で扱うために適当な大きさを持つグラフを出力することに基ずき決定した。もちろん、解の幅を広げれば正解率が高くなり、狭くすれば低くなるのであるが、システムとして妥当な値として採用した。なお評価対象文は朝日新聞８６年７月分の６４，６７９文である。

４－１　解グラフ生成率

６７，６７９文中

グラフを作成できた。

４７，３０２　（７３．１３％）

グラフを作成できない。

１７，２８９　（２６．７３％）

文が長すぎて解析できなかった。

８８　（０．１４％）

以上のように解グラフ生成率は、７３．１３％であった。特にこの中で、グラフに曖昧性がなく解が１つだったものは１０，４２８文（１６．１２％）であった。文が長すぎて解析できないというのは、プログラムの内部バッファ等の制限のため、解析対象文を２００文字の文までで切っているものである。２００文字以上の文というのは、箇条書きや社説など『。』を打たずに文を連なげている文がほとんどで一般の文を解析するためには特に問題のない数字であると思っている。実際今回の解析対象文のうち２００文字以上の文は、２５０文字以内には納まっている。この分布は４－３に述べる。

４－２　正解率

解を生成した文のうち無作為に２８０文を選び、その解析グラフ中に正解の含まれている文の数を調べた。

２８０字中
　正解が含まれている
　　２３３文　（８３．２％）
　正解が含まれていない
　　４７文　（１６．８％）

したがってすべての解析対象文に対する正解率は、

７３．１３％　×　８３．２％　＝６０．８％

であった。この数字は、この形態素解析システムの出力を受け取る次のシステムにとって大事な数字である。例えば、次の構文解析システムでは、どんなに性能がよくても、全人力文の４０％は正しい解が構文解析システムの入力以前に落ちているということを意味する。

４－３　文長と解析時間

文長と解析時間の関係を図７に示す。マシンはＳｕｎ３（メモリ８Ｍ）を使用している。また、図８にはそれぞれの文長における解グラフ生成率等も示す。グラフを見てわかる通り、文長と解析時間の関係はほぼ線形である。理論的にここで用いた拡張Ａ*のアルゴリズムは、最悪文長の２乗の解析時間がかかることが知られているが、現実の文ではグラフの広がりが発散的でなくいくつかの場所にグラフの節目ができて解析が発散的にならないために関数が線形であったということが考えれる。また解グラフ生成率のグラフからは、短い文ほど著しく解グラフ生成率が高いということがわかる。

また、１文の平均文字数は、３８．５０文字であった。

４－４　形態素の出現頻度

上述したように形態素の出現頻度を調べた。形態素の出現分布としては、Ｚｉｐｆの法則として、「形態素の出現順位がｉ番目の形態素の出現数は１／ｉに比例する」というものが知られている。今回の出現頻度の調査は、自動的に形態素解析して求めた形態素の出現分布がＺｉｐｆの法則に従うかどうかがわかる。形態素の出現頻度の上位をいくつか示す。上位は、スペースや句読点、数字等の記号や格助詞等の付属語が占めている。自立語は、「言う」「ある」等の基本的動詞が上位にある。また、選挙関係の語が多いこと等この月の新聞記事の特徴がこの結果から読み取ることができる。グラフには、出現分布として形態素の出現ランクと累計出現数の関係のグラフを示す。このグラフより出現分布は、ほぼＺｉｐｆの法則に従っていることがわかる。

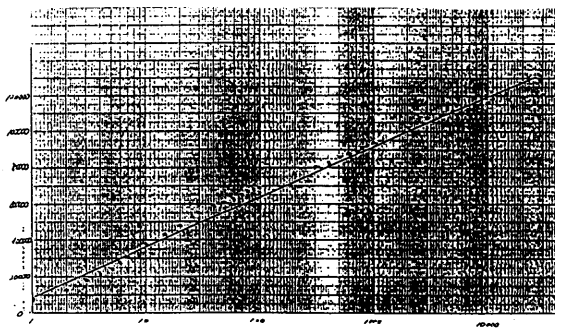

図６　形態素の出現ランクと累計出現数関係

４－４－１　動詞に関する調査

上記のように形態素の出現頻度がもとまったが、そのうち動詞について細かく調べてみた。特に動詞の辞書として有名なＩＰＡＬ辞書の見出しとの比較を行った。その結果は以下のとおりである。

	種類数	マッチ数	累計数（割合）
全和語動詞	2,223	---	29,015
IPAL和語動詞	1,476	809	21,793(75.1%)
全サ変動詞	1,843	---	9,506

表１　動詞の出現頻度

文　長	解析時間	解析成功	解析失敗	文の数	解析成功率
1 ～ 9	0.0516	2911	193	3104	0.938
10 ～ 19	0.1564	8312	1344	9656	0.861
20 ～ 29	0.3127	8949	2659	11608	0.771
30 ～ 39	0.4175	9313	3012	12325	0.756
40 ～ 49	0.6735	5356	2060	7416	0.722
50 ～ 59	0.8484	4085	1810	5895	0.693
60 ～ 69	1.0275	2865	1582	4447	0.644
70 ～ 79	1.2109	1912	1235	3147	0.608
80 ～ 89	1.3898	1289	969	2258	0.571
90 ～ 100	1.6098	884	685	1569	0.563
101 ～ 109	1.7808	525	472	997	0.527
110 ～ 119	2.0289	336	378	714	0.471
120 ～ 129	2.1345	188	291	479	0.392
130 ～ 139	2.3403	157	199	356	0.441
140 ～ 149	2.6386	73	136	209	0.349
150 ～ 159	2.7242	61	78	139	0.439
160 ～ 169	2.9594	39	64	103	0.379
170 ～ 179	2.8723	22	41	63	0.349
180 ～ 189	3.5458	18	50	68	0.265
190 ～ 200	3.9044	7	31	34	0.206
201 ～ 209	—	—	—	22	——
210 ～ 219	—	—	—	15	——
220 ～ 229	—	—	—	10	——
230 ～ 239	—	—	—	8	——
240 ～ 249	—	—	—	33	——

表２　文長と解析時間及び文の解との関係

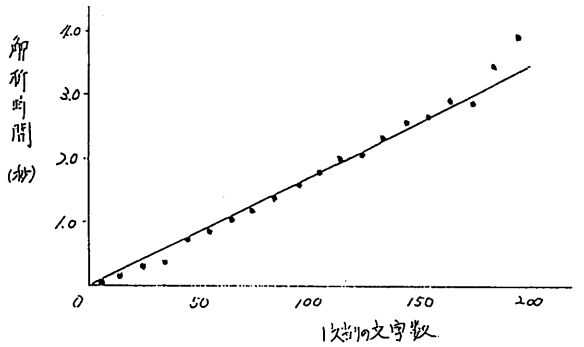

図７　文長解析時間の関係

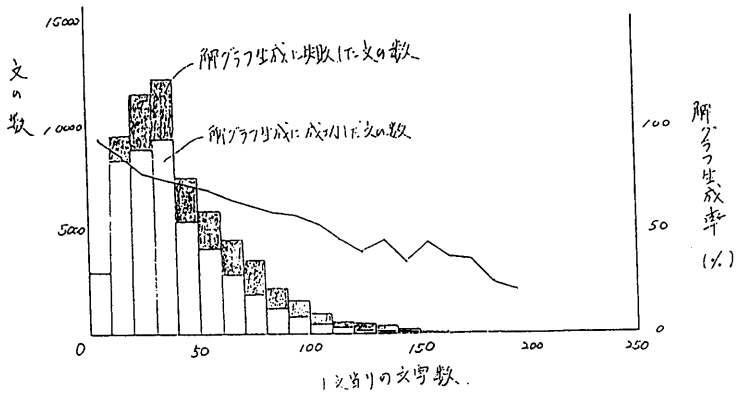

図８　文長と文の数及び解グラフ生成率との関係

（ 1274 100000 ） 73836
、（ 615 100000 ） 58497
の（ 197 100000 ） 52537
。（ 616 100000 ） 37182
に（ 196 100000 ） 31449
は（ 226 100000 ） 30127
を（ 201 100000 ） 29945
が（ 192 100000 ） 21899
と（ 195 100000 ） 19112
で（ 194 100000 ） 15896
1（ 6693 46 ） 15380
た（ 59 100000 ） 14915
0（ 65586 46 ） 12313
て（ 237 100000 ） 12230
2（ 87776 46 ） 10703
し（ 517 100000 ） 10452
」（ 838 100000 ） 9858
も（ 227 100000 ） 9826
「（ 837 100000 ） 9751
）（ 828 100000 ） 9375
（（ 827 100000 ） 9367
た（ 60 100000 ） 8296
3（ 45993 46 ） 8169
5（ 42261 46 ） 7937
4（ 47143 46 ） 6585
6（ 129344 46 ） 5359
ている（ 610 100000 ） 5004
っ（ 454 100000 ） 4825
7（ 48897 46 ） 4670
こと（ 3 100000 ） 4638
8（ 93230 46 ） 4528
から（ 193 100000 ） 4514
言（ 4717 1 ） 4101
9（ 29170 46 ） 3884
だ（ 69 100000 ） 3839
など（ 213 100000 ） 3709
の（ 253 100000 ） 3638
れ（ 186 100000 ） 3537
る（ 455 100000 ） 3290
か（ 202 100000 ） 3273
が（ 228 100000 ） 3153
し（ 47144 32 ） 3127
．（ 809 100000 ） 3021
ない（ 109 100000 ） 2925
・（ 810 100000 ） 2901
年（ 617 100000 ） 2830
し（ 402 100000 ） 2691
人（ 631 100000 ） 2673
ろ（ 456 100000 ） 2556
い（ 388 100000 ） 2549
さ（ 514 100000 ） 2477
在（ 6 7 ） 2452
な（ 511 100000 ） 2342
する（ 518 100000 ） 2319
う（ 378 100000 ） 2299
う（ 377 100000 ） 2295
っ（ 376 100000 ） 2289
，（ 808 100000 ） 2259
ずる（ 519 100000 ） 2245
日（ 664 100000 ） 2073
日本（ 88885 42 ） 2016
てい（ 609 100000 ） 1919
ない（ 110 100000 ） 1913
覚（ 80313 41 ） 1911
一（ 815 100000 ） 1876
日（ 619 100000 ） 1818
ている（ 611 100000 ） 1801
なっ（ 322 100000 ） 1794
円（ 641 100000 ） 1729
り（ 453 100000 ） 1711
な（ 70 100000 ） 1703
氏（ 1108 100000 ） 1682
と（ 239 100000 ） 1620
と（ 606 100000 ） 1559
い（ 503 100000 ） 1517
へ（ 198 100000 ） 1503
首相（ 51330 41 ） 1502
ろ（ 496 100000 ） 1501
％（ 643 100000 ） 1497
まで（ 219 100000 ） 1468
に（ 508 100000 ） 1443
就（ 76106 2 ） 1431
く（ 502 100000 ） 1427
ば（ 245 100000 ） 1426
や（ 246 100000 ） 1385
や（ 199 100000 ） 1363
もの（ 11 100000 ） 1324
ろ（ 495 100000 ） 1298
だっ（ 68 100000 ） 1247
問題（ 119306 41 ） 1233
ため（ 4 100000 ） 1209
わ（ 373 100000 ） 1199
さん（ 1107 100000 ） 1186
い（ 504 100000 ） 1159
約（ 805 100000 ） 1156

図９　出現頻度の高い形態素

4－5　解析失敗の原因

解析に失敗したということは、解のグラフを生成できなかったということである。解析に失敗した文のうち２０００文についてどのような種類の単語で解析に失敗したかということを調べてみた。以下の割合は解析失敗の全体の数に対するそれぞれの原因の数の割合である。

・略語（１１．１％）
　新聞記事特有の略語等も多い
　＜衆参、拡販、財テク、南ア、農大、口事件＞

・地名（１３．４％）
　正式名称でない国名や地域名、無名の地名
　＜フランス、ベトナム、近畿、モントルー＞

・人名（１１．３％）
　外人の名前、特殊な名前
　＜マルルーニ、ツベトフ、陳、昭如、時芳＞

・その他の固有名詞（３．８％）
　会社名、製品名等
　＜広洋、バム、ハチ公、マグナマイト＞

・一般語のカタカナ表記（１０．８％）
　普通漢字等で書くものをカタカナで書く
　＜ムリ、ムダ、ボク、ワタシ、オラア＞

・接辞（６．０％）
　主に接尾語
　＜ら、展、増、減、屑、歴、妃＞

・おどり字（々）（３．４％）
　辞書の表記はおどり字を用いていない
　＜我々、人々、日々、早々、準々決勝＞

・一般語（２９．７％）
　送りかなのゆれ等も含まれる
　＜ゆっくり、続投、取り引き、組み換え＞

・特殊なもの（１０．８％）
　上記に分類できず特殊なもの
　＜ノ（住所）、埼　玉（単語中のスペース）＞

それぞれ特に大きな理由となるものはなく、割合として小さな辞書の不備が数多く存在していることがわかる。このうち接辞やおどり字、特殊なものの一部については、すぐにでも辞書に追加することができる。しかし、地名、人名、その他の固有名詞、一般語については、特殊な地名辞典、人名辞典又は大きな辞書との突き合わせにより補充していくことが考えられる。略語は、次々と生まれてくることが考えられるため辞書に載せるだけではなく他の手段を考えねばならない。また、一般語のカタカナ表記については、カタカナ表記をしやすい単語というものが考えられるため、それを拾って辞書に追加するとよいであろう。カタカナ語はその表記に特徴があるため未知語解析が比較的容易なため、辞書に登録されてなくてもある程度解析することができるかもしれない。それぞれ、失敗の原因と、その改良の方法と見込みを以下にまとめる。

種　類	割　合	向上見込み	方　法
略語	１１．１	０	
地名	１３．４	６－８	短単位で登録
人名	１１．３	２－４	カタカナ利用、有名人
その他の固有名詞	３．８	０	
カタカナ語	１０．８	４－５	文字面により抽出、追加
接辞	６．０	５	付属語辞書に追加
おどり字	３．４	３	洗い出し、追加
一般語	２９．７	０	
特殊なもの	１０．８	３	容易なもののみ追加
	１００．０	２３－３８	

表３　グラフ作成失敗の原因とその改良の方法と見込み

4－6　誤答の原因

解グラフは生成できたがそのグラフ中に正解がなかったというものについて調査した。ここでは、解析に成功したもののうち無作為に２８０文を選びそれを人手によって調べた。そして、そのうち正しい解が含まれていなかった４７文についてその誤答の原因を調べた。調査した対象文は少ないが、ほぼ以下に分類した３つのパターンしかなかった。

誤答の原因とその割合は以下のようであった。

・辞書にないが、誤って他の語の合成等で解析できたもの　（４３．４％）
　＜靖国、小宮山、日経金、船川水産、大卒＞

・辞書にはあるが、コストの関係で他の誤った解釈をしたもの　（３９．５％）
　＜若い力、その時期、６万人、２つつけた＞

・接続の不備による他の解釈　（１７．１％）
　本来接続可能なものが接続表に接続不可能となっていたもの

このうち辞書にないものは、前節で述べたものと同様に扱われるが傾向として、１文字で意味をなす漢字の連続により単語を作成したものが多い。また略語であるものも多い。コストの関係で他の解釈が求まったものは、正しい解は辞書にあるが、正しい解の方がコストが高かったというものである。今回の評価は、一番厳しい解の幅を採用したためこの解の幅を広げれば解のグラフの中に正解は含まれる。また本質的な問題として、各品詞等につけたコストの付け方がヒューリスティックに基づいた経験的な方法だったために、このコストの付け方が最良でないことも考えられる。そのためにはシステマティックにコストの値を決定する方法を考えそれによるコストを付けなければならないだろう。接続の不備については、それらの接続を接続可能にする方向で改良すべきである。

５．まとめ

以上述べたように形態素解析システムの構築と評価を行った。結果的に約６０％の正解率が求まったが、評価の対象が新聞記事であり、結果に示したように、固有名詞や特殊な表記法などがあり、その内容を見ると妥当なところだと思う。しかし、実際に他の種類の文を対象に評価を行い有効性を確かめるべきであろう。また解析の実時間は、Ｓｕｎ　３－６０、ＭＥＭ＝８Ｍで　３０～５０文字の文が、約１～２秒で解析できているため、充分実用に耐え得ると思う。また、システムのサイズは主記憶上に約３ＭＢ、自立語辞書は、ファイルの大きさとして１８ＭＢである。しかし、これはデバック処理等が含まれており、チューンアップした場合システムが約２ＭＢ、自立語辞書は６．５ＭＢ程度になることがわかっている。なお、プログラム言語としてはＣ言語を使用している。

また、コストの幅と正解率の関係であるが、今回はコストの幅を１０として、評価を行ったが、これらの間には相関関係がある。今回はシステム側から妥当な値として、この値を選んだがこれらの相関関係についても詳しく調べる必要がある。

今後は、評価で述べた改良を行っていくと同時に言葉の持っている生産性を考え辞書になくても解析できるように、自動未知語認定について考察していく必要があるだろう。また、この形態素解析システムでは、形態素解析だけでは判断できな曖昧性をグラフで出力するという手段で解決しているが、たとえば形態素解析中に構文解析を並行に行う等、曖昧性をなるべく早い時期からなくす方法を考えたい。

６．　謝辞

本研究を進めるにあたり、九州大学辞書を提供して下さいました九州大学、ならび、ＩＰＡＬ辞書を提供して下さいましたＩＰＡ（情報処理振興事業協会）、新聞記事を提供くださった朝日新聞社に感謝いたします。

なお、本研究は筆者が松下電器産業株式会社在職中に行った研究を発表したものであります。

＜参考文献＞

1)2) 長尾・菅野
「日本語処理基本システム(1)(2)」情報処理学会全国大会第３７回　PP 1037-1040

3) 長尾・菅野
「発見的グラフ探索法を用いた日本語形態素解析」情報学会自然言語研究会　PP 74-10

4) 菅野・長尾
「曖昧さの効率的処理のための構文解析手法について」情報学会自然言語研究会　PP 74-6

5) PETER E. HART,et al.
A Formal Basis for the Heuristic Determination of Minimum Cost Paths,IEEE TRANSACTIONS ONSYSTEMS SCIENCE AND CYBERNETICS.,7,1968, PP 10-107の4)

6) Alberto Martilli
On the Complexity of Addmissible Search Algorithms,Artif.Intell.,1977,PP 1-13の5)