自然言語処理 94－4
（1993. 3. 18）

簡易日本語解析系Q_JP

亀田 雅之

(株) リコー・研究開発本部・中央研究所

概 要

多くの応用系に手軽に利用できる実用レベルの軽量の日本語解析系として、簡易日本語解析系Q_JPを試作した。形態素解析は、漢字かな混じり文での字種の特徴や和語系の複合語・派生語の特徴を利用し、小規模で簡易な辞書で解析する。構文解析では、品詞情報をベースにした構文レベルの簡単な係り受け解析を行う。また、簡易な知識の補完機能として、ユーザ介入の切り口 (対話機構) を備える。

Q_JPは、全体として簡易な系だが、比較的高い解析性能を示し、未知語によるレベルダウンがない、数十文節の長文も解析できる、といった頑強性も備えており、日本語解析系の応用の裾野の拡大や下方展開の効果が期待できる。

a Quick Japanese Parser

Masayuki KAMEDA

R & D Center, RICOH COMPANY, LTD.

16-1, Shin'ei-cho, Kohoku-ku, Yokohama, Kanagawa, 223, Japan.

ABSTRACT

This paper discusses Q_JP, a Quick Japanese Parser.

The parser features 1) simplified morphological analysis based on character types (kanji, hiragana and others), and typical characteristics of complex and derivative words in Japanese, 2) syntactic analysis of kakariuke-relationships using only morphological information, and 3) optional human interaction for disambiguation. Q_JP has the following advantages : it is fast, compact and robust, and uses a very small dictionary. These promote the development of applications using Japanese analysis.

Q_JP has been implemented as a prototype on Sun Workstation using awk.

1. はじめに

日本語の解析技術は、日英機械翻訳システムの研究開発を中心に進展を示してきたが、機械翻訳システムに限らず、日本語文書の検索・要約・理解、校正支援、日本語Ｉ/Ｆあるいは言語分析ツール等にも利用される技術であり、今後の高度な日本語処理のキーとなる重要な技術である。しかし、日本語解析のためには、数万語以上の大規模な辞書や、曖昧さや意味レベルの扱い等のために負担の大きい処理系が必要である。また、辞書については、未知語の出現やその登録作業、意味レベルの解析では、意味情報も必要になる。

こうした問題は、機械翻訳システムでは不可避の問題として対処されてきているが、一般の多くの応用系では、解析系を利用するか否かを左右する。

また、日本語解析等の研究成果が一般に広く活用されていないという指摘もある [1]。

我々は、簡易レベルではあるが、こうした利用上の障壁を緩和し、多くの応用系に手軽に利用できる実用レベルの軽量の日本語解析系の提供を目指し、簡易日本語解析系Ｑ_ＪＰのプロトタイプ版を試作した。

Ｑ_ＪＰは、字種の特徴を利用した小規模辞書ベースの形態素解析と、品詞情報に由来する構文情報による係り受け解析ベースの構文解析系からなる。辞書としては、小規模な形態素解析辞書だけの簡易な日本語解析系であるが、比較的高い解析率をもつ。

また、解析本来の曖昧さや簡易な知識の補完として、曖昧さ問合せや誤り訂正を行うユーザ介入の切り口 (対話機構) も備え、実用性能の向上を図っている。

本稿では、簡易日本語解析系Ｑ_ＪＰについて、その形態素解析と構文解析の方法について概説し、さらに、プロトタイプ版の概要と解析実験の結果を示す。

2. 形態素解析系

形態素解析では、既に実用レベルの高い解析性能が得られているが、なお、大規模辞書や未知語といった、一般の応用系での利用上の障壁が残っている。

こうした障壁を小さくするために、未知語の問題を避けながら、辞書を小規模化する。この際に、漢字かな混じり日本語文における字種の特徴に着目する。この特徴は、未知語推定、用語抽出 [2,3]、文節分割 [4]、あるいは、大規模辞書の検索を削減するための予備的な単語分割 [5,6] 等に利用されている。

ここでの形態素解析は、予備的な分割方式に近いが、これを拡張・精緻化し、また、和語系の複合語、派生語の特徴等も利用し、一般の形態素解析系のレベルに近い実用レベルの解析性能で形態素列, 単語列を導く、小規模辞書ベースの形態素解析系として独立させる。ただし、字種の特徴を利用することから、対象は漢字かな混じり日本語文に限り、漢字列からなる名詞等の複合語は分割しない、といった制限を前提にする。

2.1 小規模辞書ベースの形態素解析

日本語文の形態素解析の基本的な方式は、対象文から形態素辞書に登録された形態素を候補として切り出し、そのうち、品詞接続可能な形態素のパスを求めるものである。

ここでは、活用語は、語幹と活用語尾に分割して扱うこととし、次のような「辞書」、「字種-品詞割り当て」及び「単語合成」の分担によって、形態素候補を与える枠組みを考える。

［辞書］
- 機能語：付属語 (助詞, 助動詞, 語尾 等), 補助用言
  接続詞, 連体詞, 副詞, 形式名詞, 副詞名詞
- 少数語：サ変/カ変/上一段動詞語幹, 形容詞語幹
  「死ぬ」,「混ぜる」,「経る」 等
- 例外語：下記の品詞候補, 単語合成に外れる単語

［字種-品詞候補の割当て］
- 漢字列
  - １字　　：名詞, 五段/下一段動詞語幹
  - ２字以上：名詞, サ変名詞,
    複合動詞語幹, 形容動詞語幹*
    (*:3 字以上では接頭が「不/非/無/未」)
- カタカナ列：名詞, サ変名詞, 形容動詞語幹
- 英字列　　：名詞, サ変名詞, 形容動詞語幹
- 数字列　　：数名詞
- 記号列　　：名詞

［単語合成］
- 活用語　　：語幹＋活用語尾
- 和語複合語：動詞連用形連接の動詞, 名詞 等
- 品詞転性語：{ 名詞＋ } 動詞連用形の名詞転性 等
- 和語派生語：形容詞語幹＋派生語尾の動詞, 名詞
  動詞未然形＋派生語尾の動詞 等
- 接尾辞「的」等付加の形容動詞語幹　　他

この枠組みは、次に示すような、品詞ごとの、字種表記、語彙の数、接続上の制約の各特徴に基づく。

助詞, 助動詞 (補助用言を含む), 活用語尾

これらの付属語は、ほとんどがひらがな表記であり、語彙的には閉じた系をなす。

接続詞, 連体詞, 副詞

自立語であるが機能的な役割をもち、語彙的には、閉じた系をなす。表記上は、ひらがな表記と漢字表記が混在する。単独で文節を構成するので、接続上の制約は小さい。

名詞(サ変名詞を含む), 動詞, 形容詞, 形容動詞

概念語(内容語)と呼ばれ、語彙的には比較的多く、漢語, 外来語の名詞, 形容動詞では、開いた系をなす。単語あるいはその語幹の表記は、多くは、漢字列やカタカナ列であり、また、通常、文中では、ひらがなの付属語が後続し、接続上の制約が強い。

これらの特徴に着目し、

1. 閉じた系をなし、語彙数の少ない機能語(付属語を含む)は、辞書にもち、
2. 漢字部分は、語彙数の多い概念語(or その語幹)のいずれかであると仮定し、付属語との接続で曖昧さを解消する
3. ただし、概念語であっても、ひらがな表記やひらがなと漢字が混在する表記の語は、例外語として辞書にもつ

という基本的な枠組みを設定する。

さらに、これを次のように拡張・精緻化し、前記のような枠組みを導く。

漢字列長による字種-品詞割り当て規則の精緻化

和語の動詞, 形容詞語幹は1字漢字、形容動詞語幹は2字漢字列、その他の名詞等の漢語では2字以上の漢字列がほとんどであるという特徴に基づき、漢字の文字列長によって品詞割り当てを精緻化する。

少数語品詞の辞書登録による曖昧さの削減

動詞の細分類品詞のうち、語彙数の少ないサ変/カ変/各行の上一段他、及び形容詞は、辞書登録し、字種-品詞割り当て規則から外す。これにより、曖昧さの数を効果的に減らす。

和語系複合語/派生語の単語合成による辞書削減

和語の動詞, 形容詞は、連用形連接による複合語(「組み込む」,「切れ目」等)や派生語尾による派生語({「買う」→}「買わす」等)、連用形の名詞転性(「歩み」等)等、漢字とひらがなの混在する表記の例外語を生み出す。これらは、その形態素の並びの規則性から、「語幹+活用語尾」による単語合成と同様に扱うことで、多くの例外語を登録せずにすませる。また、「的」等の接辞による転性も同様に扱う。

字種-品詞割り当て規則の微調整

名詞連接の複合動詞(「基盤作り」等)は、見かけ上、語幹は漢字列になるので、これを候補に加える。

また、3字以上の漢字列でも、接頭辞が「不/非/無/未」であれば、形容動詞の語幹を候補に加えることで、辞書登録の削減を図る。

2.2 外部知識源による品詞の曖昧さ解消

一般の形態素解析系も含め、接続検定のレベルを超えて、構文や文脈レベルの判断が必要な、次のような曖昧さが発生することは少なくない。

・「ある」(+名詞):連体詞と動詞の連体形
・(名詞+)「で」　:格助詞と助動詞「だ」連用形
・文末の「こと」　:形式名詞と終助詞

さらに、2.1で提案した形態素解析系は、その性質上、曖昧さが残りやすい。

品詞曖昧さ解消のための規則と対話

こうした曖昧さの対処としては、曖昧のまま残す、頻度/確率が高い方を選択する、構文/文脈レベルの解析により解消する、人間に問い合わせる、といった方法が考えられる。

ここでは、曖昧さを解消するために、外部知識源として、規則と人間を考え、次のような手段を備える。

・指定レベルに応じ、特定品詞の曖昧さについて、前後の状況や頻度から、曖昧さを解消する規則
・指定に応じ、品詞の曖昧さに対して起動され、ユーザに曖昧さの解消を求める対話機構

たとえば、上記の第3例「こと」について、レベルに応じて、終助詞の頻度は少ない(あるいは扱わない)として形式名詞を選択する規則や遠方の並び(文脈)から判定する規則に従ったり、規則では解消せず、そのまま曖昧さを残す、あるいは、対話によりユーザに選択を求める、といったことが可能になる。

形容動詞語幹辞書

字種による品詞割り当てでは、2字漢字列とカタカナ列に対して、名詞, 形容動詞語幹等を与えている。しかし、形容動詞の活用語尾と助動詞「だ」の活用が、ほぼ同一であるため、これらの前の名詞と形容動詞語幹の曖昧さは解消できない。

ここでは、数の少ない形容動詞語幹を割り当て候補から外すのではなく、形容動詞語幹辞書を別途設け、曖昧さが生じた場合にだけ、参照する方式をとる。この方式で、辞書検索の節約と、形容動詞連体形「な」の場合、辞書検索なしに形容動詞語幹を検出できる。

この形容動詞語幹辞書は、[5,6]での大規模辞書を限定化したものともいえるが、ここでは、前記の規則や対話と同様に、外部知識源という位置付けで考える。

3. 構文解析系

一般の日本語構文解析では、格パターンや意味的共起といった意味レベルの情報に基づく優先解釈により、意味関係ラベル付きの解析構造を得るものが多い。しかし、これは、意味情報を必要とするため、2の小規模辞書ベースの形態素解析系と相入れない。

ここでは、系の特徴を生かすため、新たな辞書なしで、即ち、意味情報によらず、品詞レベルの情報だけに基づき、かつ簡易な構文解析系を試みる。

ただし、意味レベルの知識を持たないことの補完のために、ユーザ支援を受ける切り口(対話)を設ける。

3.1 係り受けによる簡易な構文解析の枠組み

自然言語文の解析を、句構造解析やＬＦＧで、汎用的に、厳密に扱おうとすると、細部の厳密な記述や規則の増大が生じる。また、曖昧さ(多義)の扱いでは、全解探索と適切な優先解釈による扱いがベースになるが、全解探索では、文長に対し、指数関数的に計算量(曖昧さ)が増大し、長い文を扱うのが難しい。

日本語の場合、自立語と付属語列からなる文節という単純な単位と、その後方修飾の特性に基づき、文節の係り受け構造を導く係り受け解析がある。この解析法では、文節検出は容易で、文節間のフラットな修飾対の検出が基本であるので、簡単に、直感的に実装できる。さらに、次に示す曖昧さやユーザ対話との整合性もあり、ここで求める簡易解析系に適している。

曖昧さの扱い

係り受け解析では、各文節の係り(修飾)可能な文節候補群を検出した上で、非交差制約のもとで、全体として尤もらしい各文節の係り先文節のセット(解)を求めるという組合せ最適化の枠組みとして構成することができる[7]。この枠組みは、再組合せ最適化(別解取得)が容易であり、簡易的な曖昧さの扱いができる。

また、全解型が全体の組合せを全て保持する(積算的組合せ)のに対し、部分的な可能性を保持する(加算的組合せ)だけであるので、簡易的に曖昧さを扱いながら、組合せ爆発を回避し、長文も解析可能となる。

ユーザ対話

高い解析率が望めない構文解析系にとって、また、品詞情報だけで解析するこの系では、ユーザとの対話を介した正解到達手段が非常に有用である[7]。

こうした対話では、解析結果を示したり、指示を受ける必要があるが、係り受け構造は、直感的なものとなっており、ユーザＩ/Ｆ上、有利である。

また、指示された文節の正しい修飾先を制約に、組合せ最適化(再解析)を行う方法との整合性がよい。

3.2 構文情報だけに基づく係り受け解析

前記の枠組みは、各文節の係り可能な文節対候補の検出とその候補からの優先選択の過程をもつ。ここでは、両過程を、次のように構文情報だけで行う。

係り受け可能な文節対候補の検出

係り受け可能な文節対は、「左側(係り)文節の文節末の示す修飾(係り)属性」と「右側(受け)文節の先頭の単語による被修飾(受け)属性」との組によって検出できる。これらの属性は、基本的に文節末あるいは先頭の単語の品詞や活用といった構文情報に基づくため、原則的に、意味情報は不要である。

優先解釈

格パターンや意味的な共起情報が利用されるのは、係り文節を選択する優先解釈においてである。実質的には、「係り可能な近接する文節」と「意味的に共起可能性の高い遠方の文節」との競合の選択において、その役割を果たす。しかし、距離と意味のどちらを取るべきかは微妙である。また、意味情報の設定も難しく、副作用を生じる可能性もあり、意味情報のメリットを生かすのは、技術的には難しい。

一方、日本語文の性質上、係り先は後方文節に限られ、また、構文的に修飾可能な文節は限られ、さらに、非交差制約により絞られる。また、近接する文節への係りが自然な読み方である。

こうした点を考慮すると、文節の係り先について、次のような構文レベルの簡易な経験則だけでも、ある程度のレベルの優先選択が見込める。

1. 次のようなデフォルトの優先規則とそれを補正する例外優先規則で構成する。

[デフォルト優先規則]
・文節の係り先は、
　係り可能な文節のうち、最近接文節を優先する
[例外優先規則]
・主題の「は」文節は、遠方修飾の性質がある
・読点「、」付きの文節は、最近接文節を飛越える
　：

2. 係り可能な文節が少ない程、選択の正解可能性が確率的に高くなる[7]。例えば、係り可能な文節が少ない文節から優先的に係り先を選択する戦略が有効である。当然、ユーザ対話で選択された係り受け文節対があれば、それが最優先となる。

また、上記の簡易な優先解釈の補完のためにも、ユーザ支援を受ける対話機構を設けることは、対話(ワークベンチ)形態での実用性の向上が期待できる。

4. 簡易日本語解析系Q_JP

2, 3の検討を受け、簡易日本語解析系Q_JPをワークベンチ形態 [8] のプロトタイプシステムとして、SUN SPARC Station 上にAWK言語で試作した(処理系部分 約1800行)。以下、その試作実装の概要を示す。

4.1 形態素解析系

形態素解析系は、プロトタイプでは、3つの部分ー形態素解析主処理部, 品詞曖昧さ解消部, 単語生成部ーからなる (図1)。

辞書としては、基本辞書, ひらがな辞書 (2種類)、接辞テーブルと形容動詞語幹辞書をもつ (表1)。合計で、約2800エントリ,57KBバイトと極めて小さい。

主処理部

主処理部は、基本的には、一般の形態素解析辞書、接続表を用いた処理系と同じであるが、辞書検索に失敗した場合の処理が異なる。ここでは、軽量化も考慮し、簡単な最長一致をベースにしている。

主処理部では、基本辞書による最長一致検索と接続検定を繰り返すが、検索に失敗した場合に、同一字種の文字列を切り出し、次のような処理を施す。

(1) ひらがな文字列に対しては、
　ひらがな辞書による最長一致と
　接続表による接続検定を繰り返す
(2) その他の字種は、
　接辞テーブルによる接辞処理後、
　字種-品詞割り当て規則に基づき、
　品詞候補を割当て、接続検定を行う

品詞曖昧さ解消部

主処理で残った品詞の曖昧さを解消するために、品詞曖昧さ解消規則と品詞曖昧さ解消対話を起動する。

品詞曖昧さ解消規則には、名詞と形容動詞の曖昧さがあった場合に、形容動詞語幹辞書を参照して曖昧さを解消する規則も含む。

単語合成部

単語合成部では、文節頭、単語頭、活用語尾等を品詞により認識しながら、単語合成規則により、形態素列から単語に組み上げる処理を行う。

単語合成規則には、和語系の複合語, 派生語の合成等も含む。活用語は、その終止形も併せて求める。

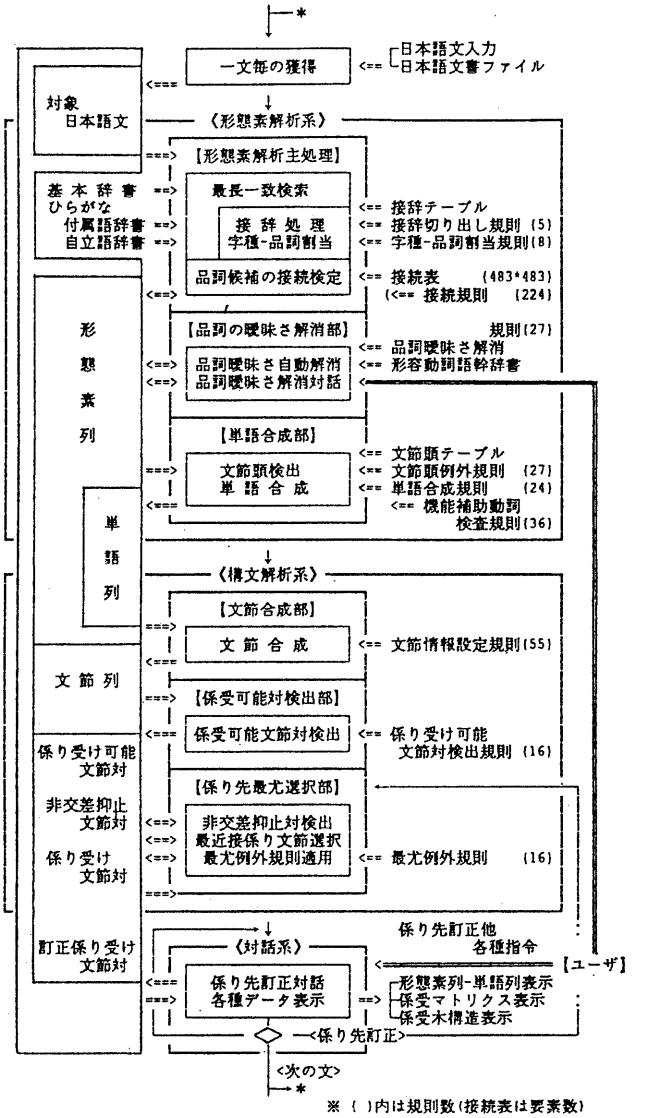

図1 簡易日本語解析系Q_JPの構成

表1 形態素解析系の辞書

辞書	エントリ	延数	サイズ	備考
基本辞書	1326	1625	32.6KB	ひらがな以外を含む例外語, 少数語, 機能語
ひらがな付属語辞書	137	417	7.4KB	ひらがな表記の助詞, 助動詞, 活用語尾等
ひらがな自立語辞書	190	256	8.1KB	ひらがな表記の自立語
接辞テーブル	12	68	0.8KB	接頭辞, 接尾辞, 末尾切出し語
形容動詞語幹辞書	1129	1129	7.8KB	漢字2字, カタカナの形容動詞語幹
合計	2794	3495	56.7KB	

4.2 構文解析系

構文解析系は、プロトタイプでは、3つの部分ー文節合成部,係り受け可能文節対検出部,係り先最尤選択部ーからなる (図1)。

文節合成部

文節合成部では、形態素解析系の形態素列 (あるいは単語列) を文節ごとにまとめながら、文節属性設定規則により、文節属性 (被修飾属性10, 修飾属性23, 補助属性11；例:図4の"→"直前の{ }内) を設定する。

係り受け可能文節対検出部

係り受け可能文節対の検出は、各文節とその後方の文節の組のうち、同検出規則により、係り受け可能な文節の組を検出する。

係り先最尤選択部

係り先最尤選択部では、構造的に係り可能な文節数が少ない文末側文節から係り先を決めていく [9]。

係り先の選択は、非交差制約を検査しながら、係り可能な最近接の文節をデフォルトで最優先にとり、以降、遠方の係り可能文節について、例外規則により必要に応じて、係り先を変更するという処理を行う。

4.3 対話機構

ワークベンチ形態の対話環境での利用形態を生かすために、ユーザの支援を受け入れる対話機能と解析結果の各種表示・出力機能がある。

対話機能としては、形態素解析の品詞の曖昧さ解消を求める対話 (4.1) と構文解析の係り受けの誤りを訂正する対話がある。後者では、訂正指示を受けると、指示された係り受け対を最優先に、他の文節の係り受けを再解析する。

4.4 解析例

形態素解析系の出力として、図2に、形態素列と単語列を示す。形態素列において、漢字表記の形態素のうち、「間」(形式名詞) 以外は、辞書に登録されていない。単語列では、活用語は、「置か」(置く=五カ未a) に見られるように、終止形を得ているだけでなく、「切れ目」(名詞) といった複合語も抽出している。

図3に、構文解析系の結果として、表示された係り受け木構造とその時の係り受け可能文節対の選択状況を示すマトリクス図を示す。この例文では、最近接修飾だけで正解が得られている。また、図4は、係り受け文節対に関する構文情報の出力である。

図5に、係り受けの訂正対話の例を示す。

例文1：日本語のように単語間に切れ目を置かない膠着言語の文の処理において、形態素解析は第一の関門である。

```
*  0[ 1):日本語      [漢字列] 名詞          1:日本語      名詞
.  6[ 2):の           ノ=格助               0:の          ノ=格助
.  8[ 3):よう         ヨウダ=ダ幹           0:ように      ヨウダ=ダ用b
|  12[ 4):に          ダ用b                 1:単語        名詞
                                            0:間          形式名詞
*  14[ 5):単語       [漢字列] 名詞          0:に          ニ=格助
.  18[ 6):間         [拡張辞書] 形式名詞    1:切れ目      名詞
.  20[ 7):に         ニ=格助               0:を          ヲ=格助
                                            1:置か        置く=五カ未a
*  22[ 8):切         [一字漢字] 下ラ基      0:ない        ナイ=ク体
|  24[ 9):れ         下ラ用                1:膠着言語    名詞
+  26[10):目         [一字漢字] 名詞        0:の          ノ=格助
.  28[11):を         ヲ=格助               1:文          名詞
                                            0:の          ノ=格助
*  30[12):置         [一字漢字] 五カ幹      1:処理        名詞
|  32[13):か         五カ未a               0:に          ニ=格助
.  34[14):な         ナイ=ク幹             0:おい        オク=五カ用b
|  36[15):い         ク体                  0:て          テ=接助
                                            0:、          読点
*  38[16):膠着言語   [漢字列] 名詞          1:形態素解析  名詞
.  46[17):の         ノ=格助               0:は          ハ=係助
                                            0:、          読点
*  48[18):文         [一字漢字] 名詞        1:第          名詞
.  50[19):の         ノ=格助               0:一          数名詞
                                            0:の          ノ=格助
*  52[20):処理       [漢字列] 名詞          1:関門        名詞
.  56[21):に         ニ=格助               1:で          ダ=用c
.  58[22):お         オク=五カ幹           0:ある        アル=五ラ終
|  60[23):い         五カ用b               0:。          句点
.  62[24):て         テ=接助
.  64[25):、         読点

*  66[26):形態素解析 [漢字列] 名詞
.  76[27):は         ハ=係助
.  78[28):、         読点

*  80[29):第         [一字漢字] 名詞
.  82[30):一         [一字数字] 数名詞
.  84[31):の         ノ=格助

*  86[32):関門       [漢字列] 名詞

*  90[33):で         ダ=用c
.  92[34):あ         アル=五ラ幹
|  94[35):る         五ラ終
.  96[36):。         句点
```

図2 例文1の形態素解析結果：形態素列(左)と単語列(右)

```
11[11)( 5) :である。                    11:である。
10[10)( 4) :├関門                       10:●関門                係り受け可能文節対
 9[ 9)( 3) :│ └第一の                    9:  ●第一の              ├・：非交差制約対
 8[ 8)( 4) :├形態素解析は、               8:●    形態素解析は、   ├○：非係先選択対
 7[ 7)( 4) :└処理において、               7:●      処理において、 └●：優先選択対
 6[ 6)( 3) :  └文の                      6:  ・・・●文の
 5[ 5)( 2)>:    └膠着言語の              5:  ・・・○●膠着言語の
 4[ 4)( 1)>:      └置かない              4:  ・・・○○●置かない
 3[ 3)( 0) :        ├切れ目を            3:              ●切れ目を
 2[ 2)( 0) :        ├単語間に            2:            ●  単語間に
 1[ 1)( 0)>:        └日本語のように      1:○           ●    日本語のように
                                           1110 9 8 7 6 5 4 3 2 1
```

図3 例文1の構文解析結果1：係り受け木構造(左)と係り受け文節マトリクス(右)

```
[1] [日本語{名詞}の{ノ=格助}ように{ヨウダ=ダ用b}] {体言句 連用}→[4]
[2] [単語{名詞}間{形式名詞}に{ニ=格助}] {体言句 に連用 終点}→[4]
[3] [切れ目{名詞}を{ヲ=格助}] {体言句 を連用}→[4]
[4] [置か{置く=五カ未a}ない{ナイ=ク体}] {動詞句 連体形連体}→[5]
[5] [膠着言語{名詞}の{ノ=格助}] {体言句 の連体}→[6]
[6] [文{名詞}の{ノ=格助}] {体言句 の連体}→[7]
[7] [処理{名詞}に{ニ=格助}おい{オク=五カ用b}て{テ=接助}、{読点}] {体言句 連用 読点}→[11]
[8] [形態素解析{名詞}は{ハ=係助}、{読点}] {体言句 連用 は 主格主題 読点}→[11]
[9] [第{名詞}一{数名詞}の{ノ=格助}] {体言句 数 の連体}→[10]
[10] [関門{名詞}] {体言句 連語連体}→[11]
[11] [で{ダ=用c}ある{アル=五ラ終}。{句点}] {純コピュラ句 用言終止}→[]
```

図4 例文1の構文解析結果2：係り受け文節対の情報

例文２：翻訳すべき言語からなる文を単語に分割し、各単語に対応する概念記号と、各概念記号間の関係を示す意味関係記号から概念構造を定め、その概念構造に基づいて翻訳すべき言語を目標言語に翻訳する翻訳システムにおいて、利用者が選択する概念記号及び意味関係記号を受け入れる手段を設け、既に選択された概念記号と意味関係記号に基づき、翻訳すべき言語からなる文の概念構造を定め、その概念構造から他言語の文を生成することを特徴とする機械翻訳方式

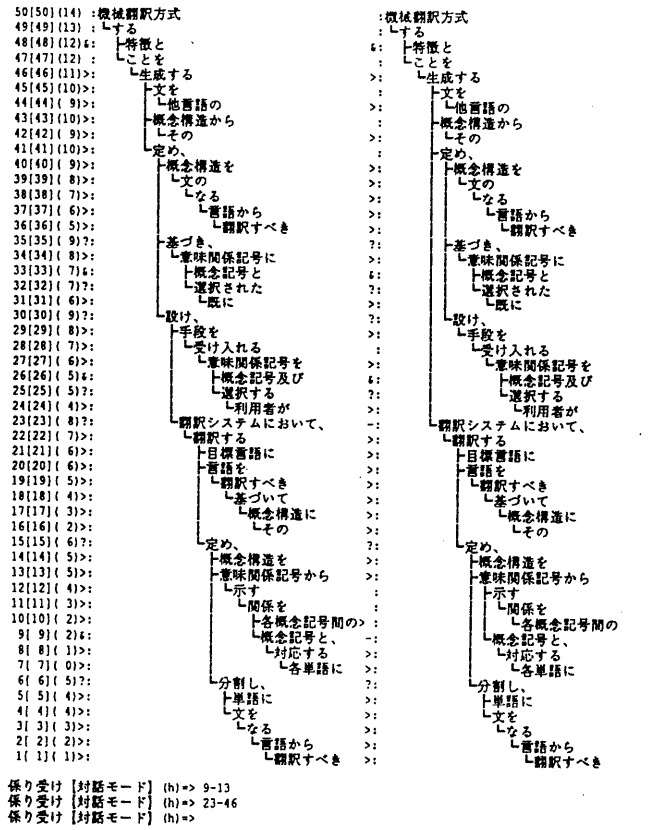

図5 例文２での係り受け訂正対話［左：訂正前,右：訂正後］

表2 形態素解析系及び構文解析系の解析正解率

	［Tuned：コーパス１］					［Blind：コーパス２］				
【文統計】	総文数	形態素数	単語数	文節数		総文数	形態素数	単語数	文節数	
	241	3~29.3~121	3~24.1~108	1~10.9~50		210	3~35.8~192	3~29.5~170	1~12.8~73	
【形態素解析系】	<0>	<1>	<2>	<3>	... <N>	<0>	<1>	<2>	<3>	... <N>
形態素正解率	99.3	99.6	99.6	99.6	... 99.6(%)	95.7	95.9	95.9	96.1	... 96.1(%)
単語正解率	99.3				... 99.6(%)	96.1				... 96.4(%)
一文正解率	87.1	92.9	92.9	92.9	... 92.9(%)	50.0	54.3	54.8	56.2	... 56.2(%)
平均対話数	0.058					0.095				
【構文解析系】	<0>	<1>	<2>	<3>	... <N>	<0>	<1>	<2>	<3>	... <N>
係り先正解率	95.1	96.5	97.4	97.5	... 97.7(%)	90.5	92.0	92.8	93.5	... 93.9(%)
一文正解率	71.0	84.2	88.0	88.4	... 88.4(%)	43.8	61.4	65.7	68.6	... 68.6(%)
平均訂正数	0.228					0.390				

注）※形態素/単語/係り先/一文正解率：形態素/単語/文節の係り先/一文全体を各単位とする正解率
※<0>～<N>：形態素解析での曖昧さ解消対話/係り受け解析での係り先訂正対話の数

表3 構文解析系の解析正解率［3～15文節］

	［Tuned：コーパス１］					［Blind：コーパス２］				
【文統計】	総文数	形態素数	単語数	文節数		総文数	形態素数	単語数	文節数	
	181	5~23.3~59	5~19.0~50	3~ 8.5~15		78	5~22.1~42	4~18.5~36	3~ 8.0~15	
【形態素解析系】	<0>	<1>	<2>	<3>	... <N>	<0>	<1>	<2>	<3>	... <N>
形態素正解率	99.6	100.0	100.0	100.0	...100.0(%)	99.4	99.7	99.8	100.0	...100.0(%)
【構文解析系】	<0>	<1>	<2>	<3>	... <N>	<0>	<1>	<2>	<3>	... <N>
係り先正解率	97.3	98.8	99.1	99.1	... 99.1(%)	93.6	96.0	97.1	98.2	... 98.2(%)
一文正解率	82.9	93.4	94.5	94.5	... 93.4(%)	70.5	87.2	91.0	93.6	... 93.6(%)
平均訂正数	0.116					0.321				

5. 解析実験

プロトタイプシステムで、解析実験を行い、解析性能及び解析速度を得た。

解析実験は、新聞,雑誌,論文,専門書,パンフレット,機械翻訳評価用例文などの約20数文献から数～10文程度づつ約200文を収集した日本語文コーパス１(チューニング用)及び同２(ブラインドテスト用)を用いた。尚、品詞曖昧さ解消規則がすべて適用されるレベル"2"に指定した。

コーパス１では、テスト解析の結果に基づき、辞書・規則及び処理系の整備・改良を施し、その最終版でのコーパス１及び同２の解析実験により、解析性能と解析速度を求めた。

5.1 解析性能

解析性能としては、形態素解析系では、形態素,単語及び１文を単位とした各正解率、構文解析系では、文節の係り先及び１文全体の係り受け構造を単位とした各正解率を示す(表2,表3)。また、対話数(N)ごとの正解率も併せて示す(単語正解率を除く)。

表２によると、ブラインドテストでは、形態素正解率(<0>:95.7%)は、[5]の単一正答率(73%)より高いが、未だ充分な正解率に達していない。構文解析系の正解率が悪いのは、コーパス１より平均文長が２割程長いせいもあるが、原因の多くは、ひらがな自立語辞書及び一部基本辞書の未整備にある。

本解析系では、辞書の未整備がほぼない状態を目標としており、その意味で、コーパス１での辞書・規則の整備状況が理想状態に近い。従って、コーパス１の解析性能をほぼ目標性能レベルと見ると、対話なしでの正解率が、形態素99%、係り先95%、１文の係り受け70%と、高いレベルが見込める。１文の係り受け正解率が、対話なしでもかなり高い。さらに、ユーザから係り受けの訂正を受けることで、１文の係り受け正解率は90%近くに達する。

表２の構文解析系の結果は、解析誤りを含んだ形態素解析結果を受けたものであり、また、かなりの長文も含んでいる。表３には、形態素解析の誤りを含まず、一般によく見られる15文節程度までの文についての構文解析系の結果を示す。両コーパスの結果を比較すると、係り受け規則も改良の余地があることがわかる。

5.2 解析速度

現在、インタプリティブに動作するＡＷＫ処理系上に実装しているため高速性はないが、解析時間は、形態素解析で、1形態素あたり約0.2秒で文長に比例し、構文解析で、およそ $(0.03\cdot(\text{文節数}-1)^2+0.09)$ 秒と、文節対の数の二乗オーダである。

例文2のような数十文節の長文でも、2-3分程度で解析できる。

6. まとめ

一般の漢字かな混じり日本語文を対象に、その字種や和語系の複合語等の特徴を利用して、小規模で簡易な形態素解析辞書による形態素解析方式を検討し、その方式に基づく試作・実験により、一般の形態素解析系に遜色のないレベルが得られることを確認し、小規模で簡易な辞書で、実用レベルの形態素解析系を実現できることを示した。

さらに、この形態素解析系の特徴を生かすため、品詞情報に由来する構文情報だけによる構文解析系を、係り受け解析に基づいた枠組みの上に試作し、浅いレベルの解析構造ながら、高い解析正解率を確認した。構文情報だけに基づく解析でも、高い解析正解率が得られたことは、通常、弱いと考えられている日本語文の構文的な制約が、必ずしも弱いとはいえないことを示唆する。

ただし、対象は漢字かな混じり日本語文に限り、漢字列からなる名詞等の複合語は分割しない、といった制限条件がある。最初の点については、通常の日本語文書は、ほとんど漢字かな混じり文になっていて、実用上問題はない。しかし、和語のひらがな表記は、しばしば現れるので、ひらがな表記されやすい和語を優先的に収集・登録する必要がある。第二点については、複合語を形成するだけ結合力が強いと考え、簡易レベルの形態素解析では許容し、分野を限定した際に、辞書により分割する方針を取ればよい。

上記の形態素解析系と構文解析系からなる簡易日本語解析系Ｑ_ＪＰは、従来の解析系に比べて、解析正解率は遜色なく、辞書・処理系ははるかに簡易になっている。未知語によるレベルダウンがない、特許文等に見られる長文も充分解析できる、といった頑強性も備えており、各種の応用系への適用に有用である。

Ｑ_ＪＰの実用化により、表層レベルで行っていた処理を形態素レベルに、または形態素から構文レベルに引き上げることによる機能向上、手軽に得られる解析結果を利用した新機能、応用系全体の軽量化、下位機種への展開等、を喚起することが期待できる。

現在は、未だプロトタイプ版レベルであるが、本来の目的に沿い、今後は、各種応用系に手軽に利用できるように、Ｃ言語でライブラリ関数化する予定である。この際に、高速性と一層の軽量化も目指す（メモリやディスク容量、速度の制約が、急速に小さくなっているが、こうした制約の減少は、各応用系が全体の高機能化,多機能化,下方展開等で享受するメリットであり、Ｑ_ＪＰのように、応用系の一部に組み込まれるモジュールは、尚、小さく、速いことが求められる）。

また、ひらがな自立語辞書や基本辞書に納めるべき単語の収集が不十分である。この整備及び係り受け規則の改良により、解析性能を目標性能に近付ける。

〔参考文献〕

[1] 島田正雄：汎用日本語解析系の試作;形態素解析コンパイラ・コンパイラの実現をめぐって，ｂｉｔ Vol.24, No.12, pp.1316-1326, 1992

[2] 高橋 他：マニュアルの索引と用語集の作成支援, 情報処理学会 第37回全国大会, 1988

[3] 東田正信：日本語記述上の形態的特徴に基づく用語抽出方式, 自然言語処理の新しい応用シンポジウム：ポジションペーパーセッション「自然言語処理の新しい応用」, 1992

[4] 坂本義行：文節単位の自動分割法－字種とひらがな連糸による－, 計量国語学 11-6,1978

[5] 長尾真, 辻井潤一, 山上明, 建部周二：国語辞書の記憶と日本語文の自動分割, 情報処理, Vol.19 No.6, pp.514-521, 1978

[6] 大橿仁司, 小池和弘, 金枝上敦史：字種切り法による形態素解析の一改良, 情報処理学会 第43回全国大会, 1992

[7] 亀田雅之, 石井信, 伊東秀夫：曖昧性解消のための対話機構を備えた日本語解析系, 情報処理学会 自然言語処理研究会報告 84-2, 1991

[8] Maruyama,H.,Watanabe,H. & Ogino,S.: An Interactive Japanese Parser for Machine Translation, Coling'90,1990

[9] 藤田克彦：決定的係り受け解析に関する試み, 第2回人工知能学会全国大会, pp399-401, 1988

〔例文出典〕

[例文1] 情報処理学会 第42回全国大会 予稿集, 1991

[例文2] 公開特許公報 (昭60), 1985