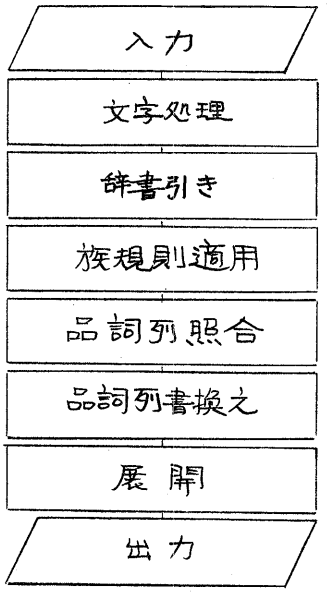
CL 2－2
(1975. 7. 25)

日本語構文解析システムやまと

西村 恕彦
(電子技術総合研究所)

1. 手続き

システムは，およそ右図に示すような手順で，日本語の文章を解析する。



(1) 入力
実験は，ローマ字表記カードを入力とした。

(2) 文字処理
あらかじめ分かち書きされている場合に，間の空白を2字ずつに調整する。

(3) 辞書引き
入力文の文字列を品詞列に書き換える。

(4) 族規則適用
品詞列の一部分を取り出し，それを構成している一つ一つの品詞について，可能ならば族を適用し，あらゆる組合せを作り出す。

(5) 品詞列照合
品詞と族からなる列を構文表と照合し，最長一致のものを定める。

(6) 品詞列書換え
構文表の指定にしたがって，品詞列を書き換える。

(7) 展開
構文解析された文の木を展開する。

(8) 出力
行印字機で印字する。

2. 実験

システムの作成と実験のうち，日本語文法の記述は，水谷静夫(東京女子大)と尾上圭介(東京大学)の作成したものを利用した。文法のコーディングと実験の実施は，渡辺芳野(IBS)による。計画全体の管理，超文法の設計，処理プログラムの作成は，西村による。

実験対象は，雑誌・新聞などからとった50個の日本語文である。特別な選択は行なっていない。表記はローマ字により，単語分かち書き30文，文節分かち書き10文，分かち書きなし10文とした。

3. 辞書引き

(1) 辞書引きは，1つの文字列を最長一致によって識別し，これを1つの品詞列で書き換える。

(2) 品詞列の長さは，通常 1 である。

けれども	←	JS	接続助詞
見込み	←	MD	一般の体言
が	←	C8	ガ

(3) 1つの文字列を長さ2以上の品詞列で書き換えることもできる。長さがゼロの品詞列（空）も許される。

△	←		空
知ら	←	D2 X1	動詞 未然
ない	←	LA X5	ナイ 連体
でき	←	D1 X2	動詞 連用
た	←	LT X5	陳述 連体

(4) ローマ字で分かち書きなしだと，次のような識別の誤りが発生した。

～の学問	←	～|のが|…
ような印象	←	よう|ない|…
～な理由	←	～|なり|…

上の例は，漢字仮名まじり表記または分かち書きによれば問題はない。しかし漢字仮名まじり表記でもうまくゆかない例は考えられる。

のようないろいろの	←	|の|よう|ない|…

(5) 活用と接続などは，文字列の分割と品詞列の分割とが，入れ子にならない場合がある。

あり	ます	ん	。
D2 X2	LH X5	..	
( )(	)(	)	

動詞 陳述 結び

どう	なっ	ている	か
FK	D2	EJ X5	J0
((	)	)(	)

句 カ

教わっ	た	こと
D2	LT X5	MD
)	))	)

動述語 句 体修語 体連語

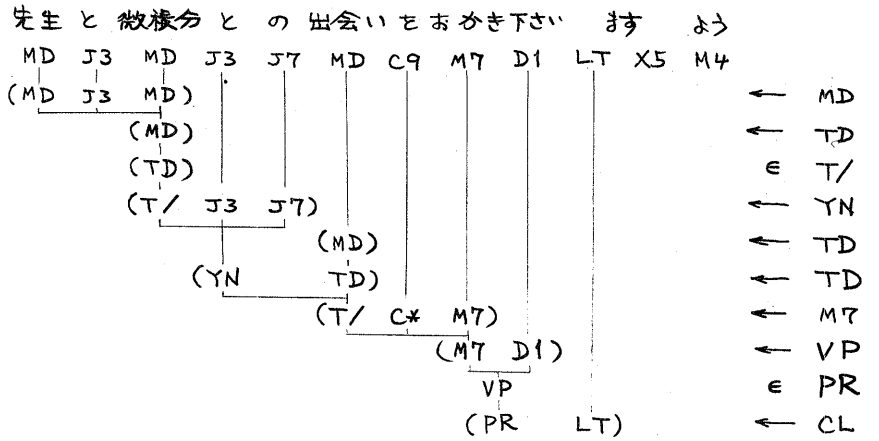
下さい	ます	よう
D1	LT X5	M4
)	))	)

動述語 句 体修語 体連語

4. 構文解析

(1) 構文解析は，1つの品詞列を1つの品詞で書き換える。品詞列の長さは通常1～6であって，その識別は最長一致による。

(2) 文脈無視の照合・書換えが大部分である。



(3) 文脈評価による照合・書換えも利用される。たとえば，

苦しかった けれども
CL X5 JS : CL (X5 JS) ← CL JS
句 連体 接続助詞

は，文脈評価形の書換えである。これをたんに，

X5 JS ← JS

と指定しても，それは適用されない。なぜならば，

CL X5 ← YN
体修語

が先に適用されてしまうからである。

(4) 品詞列は通常，品詞の値によって評価される。しかし品詞の族を適用することもできる。たとえば，次のようになる。

出会い を おかき下さい
TD C9 M7
T/ C* M7 ← M7

ここで，T/は体連語族(10品詞を代表)，C*は結合子族(9品詞を代表)である。これによって，構文規則の個数を大幅に減らすことができる。

5. 解析例

例 A

苦しかっ	た	けれども	先に	見込みが	ある	から	辛抱	でき	た	。
A3	LT X5	JS	MB	J6	MD	C8	D2 X5	C5	M7	D1 X2	LT X5	..
形容詞	陳述	接続助詞	体言	ニ	体言	ガ	動詞	カラ	サ変	動詞	陳述	結
(形述語	)(	)(	)(	)(	)(	)(	)					
句		(状況語 <	動述語 >	接続助詞	動述語	陳述						
		<	動述語 >									
		<	句 >									

< 句 >(けれども)< 句 >(から)(辛抱でき)(た) 語尾 結

< 句 >(けれども)< 動述語 >(た)
< 動述語 >< 陳述 >
< 句 > 語尾 結
文

語数 17 - 評書引き 266 ミリ秒
構文解析 12100
展開 125

例 B

私	は	何	が	どう	なっ	ている	か	知ら	ない	。
MD	G0	MD	C8	FK	D2	EJ X5	J0	D2 X1	LA X5	..
体言	ハ	体言	ガ	副詞	動詞	付加語	カ	動詞	ナイ	結
		(用修語	)	(	)(	)(	)(	)(		
		(	動述語)		カ	動述語	陳述	結		
		(<	動述語 >(ている))							
			句							

(私) ハ < 句 > カ <動述語> (ない)
(私) ハ < 動述語 > 陳述
< 動述語 > 結
< 句 > 結
文

語数 14 評書引き 207 ミリ秒
構文解析 21500
展開 100

例 C

先生	と	微積分	と	の	出会い	を	おかき	下さいます	ようお願い	いたし	ます	。
MD	J3	MD	J3	J7	MD	C9	M7	D1 LTX5	M4 M7	D1	LT X5	..

体言ト 体言 ト ノ 体言 ヲ サ変 動詞 陳述 ヨウ(体言) サ変 動詞 陳述 結

< 体言 > ( )
< 体連語 > <体連語>
< 体修語 >
< 体連語>ヲ サ変
< サ変>動詞
< 動述語>陳述
< 句 >語尾
< 体修語 >ヨウ
< 体連語 >
< 用修語>サ変
< サ変 >動詞
< 動述語>陳述
< 句 >語尾 結
文

語数 17 許書引き 235ミリ秒
構文解析 4100 ---- 特に遅い文例である。
展開 128

例 D

私	は	今日	まで	記号論理	を	教わっ	た	こと	が	あり	ません	。
MD	G0	MB	J8	MD	C9	D2	LTX5	MD	C8	D2 X2	LH X5	..

体言 ハ 体言 マデ 体言 ヲ 動詞 陳述 体言 ガ 動詞 陳述 結
(( )) (( )) (( )) ( ) ( ) ( )( )( )
体連語 (体連語マデ)(体連語 ヲ 動述語) 体連語 動述語 陳述 結
(状況語 動述語)

(私) ハ < 動述語>(た)<体連語> ガ <動述語>(ません) 結
< 動述語>陳述
< 句 >語尾
< 体修語 >体連語
(私) ハ < 体連語 >ガ"<動述語> 陳述 結
体連語 ハ < 動述語 > 陳述 結
< 動述語 > 結
< 句 > 結
文

語数 16 許書引き 238ミリ秒
構文解析 29500 ---- 特に遅い文例である。
展開 120

6. 規模

(1) 辞書

活用処理	200
付属語	250
自立語1（基本）	100
自立語2	400
計	950

(2) 品詞

語	90	語および連語の集合を類に分割したもの
連語（句）	80	
族	20	一類の集合
計	190	

(3) 構文規則

族規則	100
単文書換え規則	1000
複文書換え規則	2800
	3900

(4) 処理速度

処理速度（ミリ秒）はおよそ右のとおりである。なお，1文の平均長は24語である。構文解析の速度は，文の長さ（語数）の1.3乗ほどに比例する。したがって，短い文では0.8秒/1語，長い文では1.5秒/1語ほどになる。

	1語	1文
入力・文字処理	5	120
辞書引き	15	360
構文解析	1230	2960
展開・出力	7	180

族規則の適用がなければ，この5倍～10倍の速さになる。有限状態文法ならば，さらに5倍～10倍の速さになろう（辞書引きと同程度）。

(5) 処理結果

50文についての構文解析の結果は，ほぼ右のとおりである。複文のかかりの問題は，大部分が次のような形になった文による。

	文数
ほぼ正しい	33
複文のかかりが悪い	13
省略文が処理できない	2
その他処理できない	2

～というのだから，その役を買った専門家こそ災難だ

正 （ ）

誤 （ ）

今から振り返ってみると，関数の概念に初めて接したのは，1939年…

正 （ ）

誤 （ ）