計算言語学 21－3
（1980. 3. 14）

日本語の形態素解析について

首藤 公昭
（福岡大学・工学部）

1. まえがき

表現形式や意味内容の多様性は、自然言語の特質であり、機械処理においてこの問題を避けて通ることはできない。　筆者らは、日常の口語文章体文（技術論文等）をそのまま扱う日本語解析システムを開発すべく、基礎となる言語情報の整理を行ってきた。

量的にほう大となる体言、用言（及びこれらに相当する複合表現）に関する情報と文のわく組みに関する情報とを分離し、前者は、処理対象の分野ごとにリプレースして利用する一方、後者については、網ら的な整理を行って、分野に独立で強力な資源を与えておくことを考えている。すなわち、分野は限定されるが、取扱う文のわく組みには大きな自由度を持ったシステムの開発をねらいとしている。　本稿では、おもに後者に関する総括的な報告を行う。

日常の日本語文には、文法カテゴリーの判然としない表現や慣用句的表現がかなり頻繁に用いられている。また、日本語は、欧米語と異なり、語の屈折より膠着によって構文情報が与えられるという特徴を持つが、日常の文では、膠着の仕方はかなり繁雑で、従来、以上をまとめて体系的に整理する研究は見られない。　我々は、意味処理への見通しを重視しつつ、文のわく組みを構成する表現を網らし、文の直接構成単位と見なし得る局所的表現列の構造を明らかにした。　この局所的表現列は、「文節」の考えを意味のうえから拡張したものであり、我々は、「E-文節」と呼んでいる。　欧米語における語の屈折を取扱う段階との類似性から、E-文節の構造を解析する段階を（広義の）「形態素解析」と呼ぶこともできよう。

なお、日本語にも膠着（以後、接続という）の仕方と密接に関連した屈折、すなわち「活用」があり、我々は、これに関する機械処理向きの整理を行い、ほとんど例外なく処理できる見通しを得ている。

2. 要素表現の設定

意味機能の点で単語的に扱うのが望ましいと思われる複合表現も含めて、分野に余り関係なく使用され、量もさほどほう大とはならない表現を網ら的に収集した。　資料は、高校の教科書、技術論文の口語文章体、約18,000文である。

2.1 付属語的表現

日本語は膠着言語、あるいは後置詞言語であり、文のわく組みは、後置詞、すなわち付属語によって与えられると云われる。　しかし、意味の点から日常の文を見れば、付属語と類似の機能を有する複合表現が多数見出される。

例えば、

'それについても考えておかなければならないのではなかろうか。'

という文でいわゆる付属語を文頭側から抽出すれば、'に'、'て'、'も'、'で'、'ない'、'ば'、'ない'、'の'、'で'、'は'、'ない'、'う'、'か'となるが、この様な解析を行うのは、文の意味を把握する上で得策でない。

そこで、人間が理解する際に一まとまりに認識すると思われる'について'、'も'、'ておく'、'なければならない'、'ので

はなかろうか'等を付属語的な要素表現とみなす。この種の表現が日常、どれくらい使われるかが問題であるが我々は、前記の資料から約950種を得、その後の調査からかなり網らできたのではないかと考えている。これらの表現のうち、'について'、'における'など格助詞、副助詞的なものを「関係表現」、'なければならない'、'のではなかろうか'など助動詞、終助詞的なものを「助述表現」と呼んでいる。

広範な日本語文を扱うシステムを考える際、前者は、日本語の表層格、後者は、モーダリティ、アスペクト等を与える重要な表現群である。

2.2 接尾語的表現

日常の文には、多数の接尾語的表現が用いられるが、これらのうち、分野に余り関係なく、高い頻度で用いられるもの約120種を抽出した。例えば、'調査するかどうかが問題なのである'という文からは、'する'、'かどうか'、'なのである'を抽出した。その他の接尾語については、接頭語、助数詞等と同様、分野を定めて整理することを考えている。

2.3 自立語的表現

接続詞、連体詞、副詞的な表現の抽出を行った。ここでも'このような'、'しかしながら'、'従って'、'ある程度'などの複合表現を含め、計約400種を定めたが、必ずしも十分とは云えないため、現在、増補中である。

3. 文の局所的表現列(E-文節)の構造

3.1 要素表現の分類

2.で定めた要素表現、特に付属語、接尾語的表現の接続機能を詳細に調べ、これに基づいて分類を行って約140個の文法カテゴリーを定めた。これは、学校文法での文法カテゴリーを接続機能によってリファインしたもので、同綴り異機能(異義)の要素表現は、機能ごとに別表現とみなした分類となっており、接続の解析によって要素表現の多義選択が可能となる様、配慮している。

分類の概要を表1に示す。(表現例については図1を参照のこと)

表1 要素表現の分類

大分類			表現数	文法カテゴリー(code)
付属語的表現	関係表現	R_{NP1}	155	R01 R09,R10~R13,R17~R19,R20~R27
		R_{NP2}	65	R30~R39,R40~R46,R50~R59,R60~R63
		R_{NP3}	16	R69
		R_{NN1}	1	R70
		R_{NN2}	119	R71
		R_{NN3}	4	R72
		R_{PN}	38	R80
		R_{PP1}	150	R90~R93
		R_{PP2}	1	R94
		R_{PP3}	39	R95~R97
	助述表現	$*A_{np1}$	41	A11,A15,A18,A1D,A14,A17
		$*A_{np2}$	4	A24,A29
		$*A_{np3}$	3	A30
		$*A_{pp1}$	299	A40,A41,A43~A49,A4A~A4D
		$*A_{pp2}$	13	A50
		$*A_{pp3}$	4	A60
		$*A_{pp4}$	1	A70
接尾語的表現		$(S_{nn1}$	290	S50~S59,S60~S69,S70~S74)
		S_{nn2}	6	S80
		S_{nn3}	4	S90,S91
		S_{pn}	92	S40~S49
		$*S_{np1}$	12	S21,S29,S2A,S2D
		$*S_{np2}$	2	S35,S34
		$*S_{pp}$	2	S11,S1D
自立語的表現	接続詞	C_1	36	C10
		C_2	87	C20
	副詞	D_1	185	D10
		D_2	80	D20
	連体詞	T	164	T00
	名詞	$(M_1$	-	M11~M15)
		$(M_2$	-	M20)
	用言	Y	-	Y00~Y09,Y0A,Y0B)

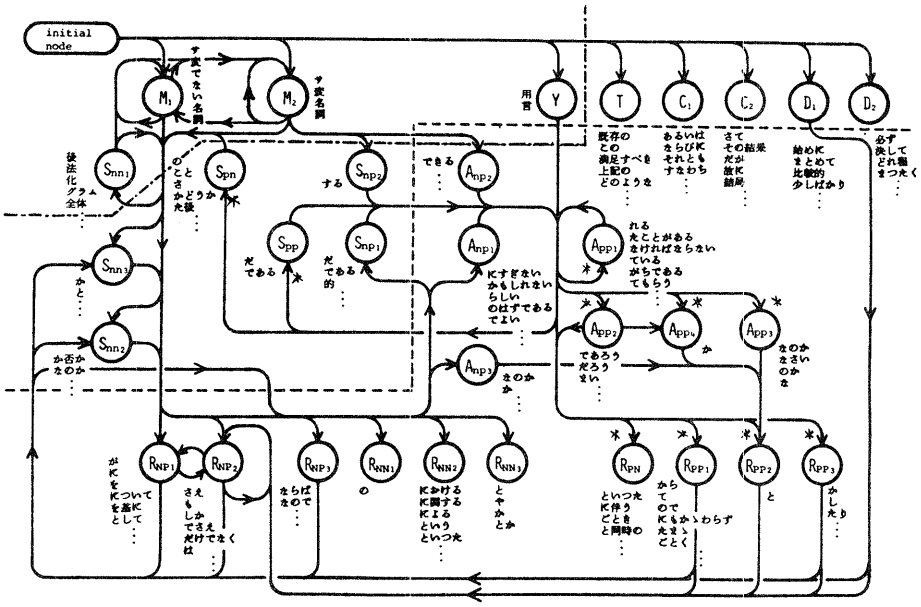

図1　接続ルールの概要

れば、不変化部'KIS'を語幹とすることができる。

(2)便役、可能動詞の語尾を切離し、独立した助述表現とみなすことができる。

例えば、便役動詞'降らす'は、'FURASU'となり、語幹'FUR'に便役の助述表現'ASU'が接続したものとする。

また、可能動詞'行ける'は、'行く'の語幹'IK'に助述表現'ERU'が接続したものとする。

(3)英字入出力機器が使える。

(4)仮名との相互変換が容易である。

(5)表記としての国際性。

筆者らは、特に、(1)、(2)を重視し、各行で子音部文字が変化しない表記法を採用した。これを表2に示す。

3.2 接続ルール

3.1で定めた各カテゴリーに対して、どのカテゴリーに接続可能かを規定するルールを与えた。ルール数は、被接続、接続カテゴリーの対、約3,600個である。ルールの詳細は略すが、その概要を表1の大分類のカテゴリー32種を用いて、遷移図の形で図1に示す。

ノードは、大分類でのカテゴリー、矢印は起点のノードの表現に着点のノードの表現が文末側から接続できることを表す。また、ノードの遷移の過程で、破線を上部の領域から下部の領域へ最後に横切る時が、E-文節の自立部と付属部の境界に大よそ対応している。

3.3 ローマ字表記

2.で抽出した各種表現は、現在、ローマ字表記によっている。

ローマ字表記には、次の利点がある。

(1)語幹、活用語尾の分離が音素レベルで行え、活用のルール化が容易になる。

例えば、サ変動詞'期する'の連用形は'期し'となるが、ローマ字表記によ

表2　採用したローマ字表記

A	I	U	E	O				
ア	イ	ウ	エ	オ				
KA	KI	KU	KE	KO	KYA	KYU		KYO
カ	キ	ク	ケ	コ	キャ	キュ		キョ
SA	SI	SU	SE	SO	SHA	SHU	SHE	SHO
サ	シ	ス	セ	ソ	シャ	シュ	シェ	ショ
TA	TI	TU	TE	TO	CHA	CHU	CHE	CHO
タ	チ	ツ	テ	ト	チャ	チュ	チェ	チョ
NA	NI	NU	NE	NO	NYA	NYU		NYO
ナ	ニ	ヌ	ネ	ノ	ニャ	ニュ		ニョ
HA	HI	HU	HE	HO	HYA	HYU		HYO
ハ	ヒ	フ	ヘ	ホ	ヒャ	ヒュ		ヒョ
MA	MI	MU	ME	MO	MYA	MYU		MYO
マ	ミ	ム	メ	モ	ミャ	ミュ		ミョ
YA		YU		YO				
ヤ		ユ		ヨ				
RA	RI	RU	RE	RO	RYA	RYU		RYO
ラ	リ	ル	レ	ロ	リャ	リュ		リョ
WA				WO				
ワ				ヲ				
Q								
ン								
GA	GI	GU	GE	GO	GYA	GYU		GYO
ガ	ギ	グ	ゲ	ゴ	ギャ	ギュ		ギョ
ZA	ZI	ZU	ZE	ZO	JA	JU	JE	JO
ザ	ジ	ズ	ゼ	ゾ	ジャ	ジュ	ジェ	ジョ
DA	DI	DU	DE	DO				
ダ	ヂ	ヅ	デ	ド				
BA	BI	BU	BE	BO	BYA	BYU		BYO
バ	ビ	ブ	ベ	ボ	ビャ	ビュ		ビョ
PA	PI	PU	PE	PO	PYA	PYU		PYO
パ	ピ	プ	ペ	ポ	ピャ	ピュ		ピョ

これは、ヘボン式と日本式を組合せたものである。ヘボン式単独の場合は、例えば、動詞'立つ'の未然、連用、終止形は、それぞれ'TATA'、'TACHI'、'TATSU'となって、語幹部が取りにくい。しかし、日本式によれば、'TATA'、'TATI'、'TATU'の不変化部'TAT'を語幹とし、語尾'A'、'I'、'U'を他の動詞，例えば'起こす'などの語尾変化のパターンと共通にすることができる。

3.4 活用の整理

上記の表記法に基づき、活用語尾、活用形、活用型を表3のようにまとめた。　表中、＊印は、音便変化した1文字を表す。

ここでの整理の特徴を列挙すれば、次の通りである。

(1) 通常の口語文章体文を対象とするため、ウ音便を除外している（また、現時点では、'です'、'ます'等の丁ねい表現も考慮していない）が、この範囲ではほとんど例外なく処理できる。

(2) 音便の整理（後述）に基づき、5段型にI，T，Q型の3種を加えた。

(3) '言う'などに特定の活用型（7型）を与え、'IWU'などの変化を避けた。

(4) 'ず'、'まい'、'う'などを非活用表現とみなし、活用処理から除外した。

表4　音便変化する文字

活用型	見出し語の末尾1文字	入力文字列中の対応する1文字	見出し語の例	入力の例
0	G	I	TUG（継ぐ）	TUIDA
	K		KIK（聞く）	KIITA
1	R	T	OKOR（起こる）	OKOTTA
	K		IK（行く）	ITTA
	T		KAT（勝つ）	KATTA
2	N	Q	SIN（死ぬ）	SIQDA
	B		YOB（呼ぶ）	YOQDA
	M		YOM（読む）	YOQDA

3.5 音便の整理

音便変化する語幹の末尾1文字を表4のように整理した。

3.6 活用に関する接続条件

表1で＊印を付したカテゴリーの表現は活用するものであり、これらに文末側から何らかの表現が接続する際（図1で＊印の接続）、特定の活用型、および活用形であることが要求される。

3.2で述べた接続ルールには、この種の制約は含まれておらず、接続ルール表と切離して、活用表現に文末側から接続可能な各表現の辞書項目として与える。

この項目は、何活用の何形に接続可能かを表3に基づいて記した

表3　活用の整理

活用形 / 活用型	Code	未然 1	未然 2	連用 3	連用 4	連用 5	終止 6	連体 7	仮定 8	仮定 9	命令 A	語幹 B
動詞型 5段音便I型	0	A	A	O	I	I*	U	U	E	E		ex. KIK(聞く)
動詞型 5段音便T型	1	A	A	O	I	T*	U	U	E	E		ex. OKOR(起こる)
動詞型 5段音便Q型	2	A	A	O	I	Q*	U	U	E	E		ex. SIN(死ぬ)
動詞型 5段型	3	A	A	O	I	I	U	U	E	E		ex. KES(消す)
動詞型 1段型	4	ε	ε	YO	ε	ε	RU	RU	RE	RO	YO	ex. TOZI(閉じる)
動詞型 サ変型	5	I	E	IYO	I	I	URU	URU	URE	IRO	EYO	S (する)
動詞型 カ変型	6	O	O	OYO	I	I	URU	URU	URE	OI		K (来る)
動詞型 W型	7	WA	WA	O	I	T	U	U	E	E		ex. TIGA(違う)
形容詞型	8	KU		KARO	KU	KAT	I	I	KERE			ex. YO (良い)
形容動詞型 NA型	9				NI			NA				ex. KIREI(きれい)
形容動詞型 NO型	A				NI	ε		NO				ex. HODO(ほど)
形容動詞型 ε型	B				NI			ε				ex. ONAZI(同じ)
た型	C	E		ARO			A	A				T (た)
だ型	D	E	EHA	ARO	E	AT	A	A				D (だ)

もので、表5に示す45種のいずれか一つをコードで与えておく。例えば、助述表現'ているところだ'に対しては、K-43が記され、2型を除く動詞型活用表現の活用形5に接続することが示される。

この条件をチェックするには、被接続の活用表現がどの活用型であるかを知らねばならないが、表1に示した活用表現の小分類は、活用型による分類となっており、コード3桁中、最下位桁は表3に示した活用型コードそのものである。例えば、用言の動詞、形容詞、形容動詞への細分類は、活用型によって行われていることになる。

表5 活用接続条件

Code	接続可能な活用型、活用形
K-01	(5 B)
K-02	(0-3 B)
K-03	(0-4 B)(7 B)
K-06	(8 B)
K-07	(A B)
K-08	(9-B B)
K-09	(8-B B)
K-11	(0-3 1)(7 1)
K-12	(0-7 1)
K-13	(0-8 1)
K-14	(0-8 1)(D 1)
K-15	(0-8 1)(D 1)(D 2)
K-21	(0-7 2)
K-22	(4 2)(6 2)
K-25	(0-7 3)
K-26	(0-8 3)
K-27	(0-8 3)(D 3)
K-31	(0-7 4)
K-33	(0-7 4)(8 B)
K-34	(0-7 4)(8-B B)
K-36	(0-B 4)
K-41	(0 5)(2 5)
K-42	(0 5)(2 5)(9-B B)
K-43	(0-1 5)(3-7 5)
K-45	(0-1 5)(3-7 5)(8 4)
K-46	(0-1 5)(3-8 5)(D 5)
K-51	(0-7 6)
K-52	(0-7 6)(C-D 6)
K-53	(0-8 6)
K-54	(0-8 6)(9-B B)
K-55	(0-8 6)(9-B B)(C-D 6)
K-56	(0-8 6)(C-D 6)
K-57	(0-8 6)(D 6)
K-61	(0-7 7)
K-62	(0-7 7)(C-D 7)
K-64	(0-8 7)(C-D 7)
K-66	(0-8 7)(9-B B)(C-D 7)
K-67	(0-9 7)
K-68	(0-9 7)(B-D 7)
K-69	(0-9 7)(C-D 7)
K-70	(0-B 7)
K-71	(0-D 7)
K-72	(8-D 7)
K-73	(C-D 7)
K-78	(0-8 8)

4. 辞書

2.で定めた要素表現を収録した機械辞書を作成した。見出し語に続く文法情報の項目は、表1の文法カテゴリーコード、および表5の活用接続条件コードの2種である。一部を図2に示す。

```
0230 00  TOHAKAGIRANA C-A48 K-55
0231 12  TOHAKAGIRAZU C-R95 K-55
0232 12  TOHIKAKUSITE C-R50
0233 12  TOIUKATATIDE C-R17
0234 12  TOIUKEITAIDE C-R17
0235 12  TOKAQKEINAKU C-R09
0236 12  TOKAQREQSITA C-R71
0237 12  TOKAQREQSITE C-R17
0238 12  TOKAQREQSURU C-R71
0239 12  TOMUKAQKEINI C-R09
0240 12  TOMUKAQKEINO C-R71
0241 12  TOONAZIKURAI C-R52
0242 12  TOUITUTEKINI C-D10
0243 12  TOHAKAGIRAZU C-R69
0244 12  TUGININOBERU C-T00
0245 12  UETODOUYOUNI C-D10
0246 12  UTINOIZUREKA C-S48 K-71
0247 12  WOMOKUTEKINI C-R08
0248 12  WOMOKUTEKINO C-R71
0249 12  WOMOTONISITA C-R71
0250 12  WOMOTONISITE C-R19
0251 12  YOUNIOMOWARE C-A44 K-71
```

図2 機械辞書の一部

5. むすび

本稿では、対象分野に余り依存しないと思われる言語情報を整理し、日常の日本語文の形態素解析を行う際に有効なルールの体系を示した。本研究のおもな特徴は、長単位の要素表現を広く取入れて構造規定を行っており、意味処理との繋がりが直接的であるこ

と、およびかなり広範な言語表現に対処し得ることである。取扱う言語表現の範囲を広げることは、それだけ適格な意味の表現・伝達が可能となることであり、文解析でも、よりあいまいさの少い意味解析の可能性をもたらす。無論、このためには、多様な表現について、使用される状況との関連も含めて、意味の整理を行っておく必要がある。我々は、助辞表現、関係表現の意味分類を行っているが、これについては別の機会に報告したい。

本稿で整理した言語情報は、資源としてまとめると、

- 機械辞書（図2参照）
- 接続ルール表（図1参照）
- 活用語尾表（表3）
- 音便変化文字表（表4）
- （・活用接続条件表（表5））

であり、我々は、すでにこれらを用いた形態素解析実験システムを作成し、要素表現の網羅性、接続ルール（活用に関するルールを含む）の妥当性に関する検証を行って良好な結果を得ており[2]、技術論文等の口語文章体文において体言、用言（および相当する表現）を除いた部分の構造は、ほとんど例外なく整理できたと考えている。

また、本稿の文法ルールは、要素表現の意味による、文中の隣接する場所での共起制限を含んでおり、接続を検定する局所レベルの解析で要素表現の多義の低減が可能な限り行える様に配慮されている点も確認ずみである。例えば、'開発したため、…'における'ため'の意味は、先行する'た'との共起制限から目的の意味を取り得ず、原因の意味に限定できる。

接頭語、接尾語を含む複合名詞等の構造解析の問題、構文・意味解析への橋渡しの問題等、残された問題は多く、現在、検討中である。また、上記の資源に字種情報を加えて、べた書き文の自動分かち書きに利用する実験を行っている。

参考文献

(1) 首藤、橋原、吉田："日本語の機械処理のための文節構造モデル"，信学論(D)，J62-D．12（1979）．

(2) 首藤公昭："文節の構造と文解析"，電気四学会連大(昭54)，(1979)．