自然言語処理 58 － 2
（1986. 11. 21）

知識ベースの生成を目的とした ニュース文の解析

遠藤 勉

大分大学工学部

科学技術情報に関する質問応答システムを作成する際に問題となるものの一つに知識ベースの構築があるが、雑誌や新聞のニュース記事はそのための重要な情報源となる。本報告では、ニュース記事から知識ベース構築のための情報を自動的に抽出することを目的とした、日本語ニュース文の解析法について述べる。対象とした言語資料は、雑誌「日経エレクトロニクス」のニュース文のうち計算機に関するものである。まず、記者が、どのような過程をへて文章表現に至ったかを分析し、ニュース文の生成モデルを提案する。次にモデルの中でニュース文の言語表現を直接規定している記者の認識構造を記述するためのメタ言語Mを定義する。Mは言語の語いの集合であるプリミティブＰ１、プリミティブ間の関係を表すパターンＰ２、パターン間の関係を表すプロダクションＰ３から構成されており、これを用いてニュース文の意味を記述する。最後に、メタ言語Mとニュース文の対応関係を与え、構文・意味解析の流れを示す。

Japanese News Story Analysis for Knowledge Base Construction

Tsutomu ENDO

Department of Information Science and Systems Engineering
Oita University
700 Dannoharu, Oita 870-11, Japan

The news story in scientific journals is a important source to extract information for the knowledge base of a question answering system. In this paper, we discuss the method of analysing the Japanese news story on a special world of computers, based on the process how a journalist wrote it. At first we present the generation model of a news story, and then introduce the meta language M which consists of primitives, patterns and productions in order to describe the model. The meaning of a news story is represented in M. Finally, we show the flow of syntactic and semantic analysis using the relation between M and the sentences in the news story.

1．まえがき

自然言語によるデータベースの照会あるいは質問応答システムの研究は国内外を通じて活発に行われているが[1]~[5]検索の対象となるべき知識ベース（データベース）の生成は人手で行なうことが多い。この種の自動化への試みが機械翻訳などに比べて少ないのは、知識源となるべき情報を主題分析（要約）し、その内容をシステムの内部表現形式に写像するという高度な処理を含むためと考えられる。しかしながら、絹川らは記事文検索のためのキーワードへのロール付与を、格文法に基づく日本語ニュース文の構造解析により自動化している[6]。また高松らは特許請求範囲文などの日本語技術抄録文から、抽出項目を指定したフレームを用いて構造情報を抽出し、関係形式などにデータベース化する手法を提案している[7]。

筆者らも、計算機の歴史を対象とした日本語による質問応答システムを作成したが[8][9]、その知識ベースは雑誌や図書などを参考にして手作業で入力したものである。ここで扱っている計算機の歴史などのような科学技術情報に関する知識ベースの構成においては、雑誌や新聞のニュース記事が重要な情報源となる。そこで、ニュース記事に表現されている内容を自動的に抽出できれば、知識ベース構築の省力化になるはずである。

本報告は、質問応答システムの知識ベースの生成を目的とした、日本語ニュース文の解析法について述べたものである。この解析法の特徴は、ニュース記者がどのような認識構造をへて文章表現に至ったかという過程に着目して、言語情報を分析・分類したことにある。なお、分析の対象とした資料は、雑誌「日経エレクトロニクス」の「ＮＥレポート」や「産業ニュースダイジェスト」の中の計算機システムの発展や動向に関するニュース文である。

次章以下では、既存の質問応答システムの概要（第２章）、ニュース文の生成モデル（第３章）、モデルを記述するためのメタ言語（第４章）、ニュース文とメタ言語表現との対応（第５章）について述べる。

2．計算機の歴史に関する質問応答

ここでは、本研究の動機を与えた日本語質問応答システムの概要を示す。図１がシステム構成である。英数字、カナよりなる質問文は形態素解析により文節単位に分割される。次に、格文法に基づく係り受け規則を用いて文節間の係り受け関係を調べ、述語を中心とした質問文の格構造を求めるとともに、抽出規則により質問のタイプを決定する。最後に、図２のような表形式の知識ベースの属性に格構造を対応づける変換規則を使用して検索用のデータ言語を生成する。図３が本システムによる質問応答の例である。

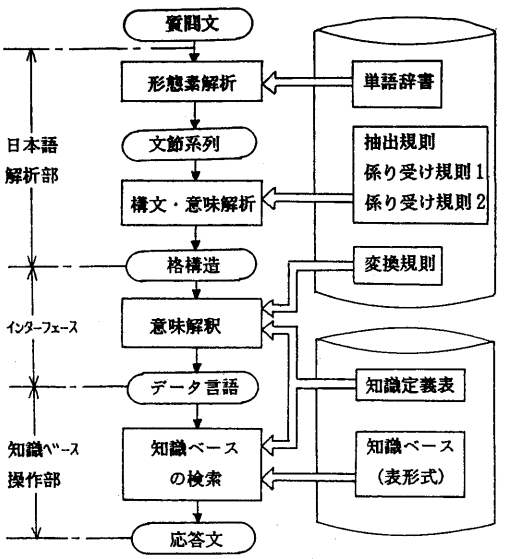

図１　質問応答システムの構成

［ＥＶＥＮＴ－Ｒ］

ENO	対象	主体	年代	場所	事象
1	ENIAC	ECKERT	1946	ｱﾒﾘｶ	作成
2	ETL-MARK3	高橋茂	1956	日本	作成
3	FUJIC	岡崎文次	1956	日本	作成
4	CRAY-1	CRAY	1976	ｱﾒﾘｶ	発表
5	4004	F.FAGGIN	1971	ｱﾒﾘｶ	発表

［ＯＢＪ－Ｒ］

対象	カテゴリ	素子	特徴	用途
ENIAC	ｺﾝﾋﾟｭｰﾀ	真空管	－	弾道計算用
ETL-MARK3	ｺﾝﾋﾟｭｰﾀ	ﾄﾗﾝｼﾞｽﾀ	ｶﾞﾗｽ遅延線ﾒﾓﾘ	－
FUJIC	ｺﾝﾋﾟｭｰﾀ	真空管	水銀遅延線ﾒﾓﾘ	ﾚﾝｽﾞ設計用
CRAY-1	ｽｰﾊﾟｰｺﾝﾋﾟｭｰﾀ	LSI	ﾊﾟｲﾌﾟﾗｲﾝ方式	高速計算用
4004	ﾏｲｸﾛｺﾝﾋﾟｭｰﾀ	MOS	4ﾋﾞｯﾄ	電卓用

図２　知識ベースの一部

ｾｶｲﾃﾞｻｲｼｮﾆﾂｸﾗﾚﾀﾏｲｸﾛｺﾝﾋﾟｭｰﾀﾊﾅﾝﾃﾞｽｶ?
世界で最初に作成されたﾏｲｸﾛｺﾝﾋﾟｭｰﾀ は4004です。
ｿﾚﾊｲﾂﾂｸﾗﾚﾏｼﾀｶ?
4004は1971年に作成されました。
ｿﾚｶﾞﾂｸﾗﾚﾀﾉﾊｲﾂﾃﾞｽｶ?
4004が作成されたのは1971年です。
ｿﾚﾊﾄﾞｺﾃﾞﾂｸﾗﾚﾏｼﾀｶ?
4004はINTEL で作成されました。
ｿﾚﾊﾄﾞﾝﾅｺﾝﾋﾟｭｰﾀﾃﾞｽｶ?
4004はMOSを使った4ﾋﾞｯﾄの電卓用のﾏｲｸﾛｺﾝﾋﾟｭｰﾀです。
ｿﾚﾊﾀﾞﾚｶﾞﾂｸｯﾀｲﾏｼﾀｶ?
4004はF.FAGGINが設計しました。
ｿﾉﾋﾄﾂｸｲｼｬﾊﾀﾞﾚﾃﾞｽｶ?
4004を設計した人はF.FAGGINです。
F.FAGGINﾊﾄﾞｺﾉﾋﾄﾃﾞｽｶ?
F.FAGGINはｱﾒﾘｶの人です。

図３　質問応答の例

3．ニュース文生成モデル

ニュース文には、計算機に関する様々な事物や事象およびそれらの関係を直接概念としてとらえた表現だけでなく、認識主体（記者）の観念的な前提、判断、期待、意図などが概念と結びついて表現されており、これらが主体の立場の移行や思考の展開に応じて１つの記事としてまとまったものとみなされる。そこで、図４に示すようなニュース文の生成モデルを考えた。簡単な例を用いて生成の流れを説明する。

（１）外界の事象として、Ｂ社がｘをＣＰＵとするパソコンＰｘをその仕様とともに発表したとする。

（２）記者の前提（知識）として次の事を仮定する。

（ｉ）Ｂ社のライバルのＡ社はすでにｘをＣＰＵとするパソコンを発表している。

（ii）ｘをＣＰＵとすればαをサポートしているはずである。

（iii）Ｂ社の開発状況のうわさをいろいろきいている。

（３）外界からの事実（１）の認識と（２）の前提が結びついた結果、次のようなニュース文が想定される。

「Ｂ社もｘをＣＰＵとするパソコンＰｘを発表した。性能は、・・・・・である。しかしαは装備していない。年内にはサポートされるもようである。」

下線部が記者の能動的な認識と概念との結びつきに対応する表現である。

4．認識構造の記述

図４のモデルから、ニュース文の言語表現を直接規定しているのは主体の認識構造であることがわかる。そこで、この構造を記述するための枠組み（一種のメタ言語）Ｍを定義する。

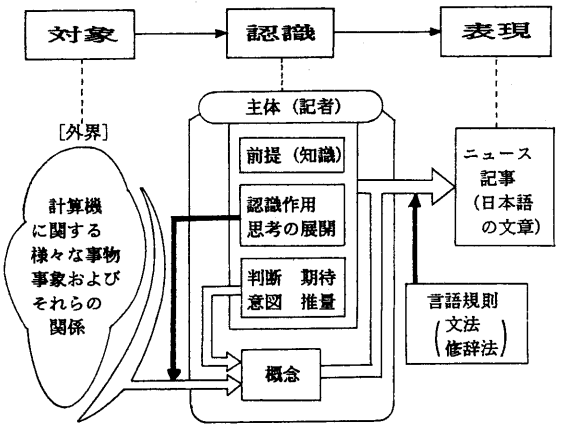

図４　ニュース文生成モデル

Ｍ＝（Ｐ１，Ｐ２，Ｐ３）

Ｐ１はメタ言語の語いの集合に相当するもので、プリミティブ（Primitives）と呼ぶ。Ｐ２はプリミティブ間の関係を表しており、パターン（Patterns）と呼ぶ。Ｐ３はパターン間の関係を生成規則としてとらえたもので、プロダクション（Productions）とよぶことにする。これらはモデルからもわかるように、本来言語表現とは独立して設定されるべきものであるが、ここでは、ニュース文の分析を通して、その抽出を行なった。

4．1　プリミティブ

プリミティブとして、Ｃ（概念プリミティブ）、Ｕ（判断プリミティブ）、Φ（認識プリミティブ）、Λ（関係プリミティブ）の４種類を設けた。

Ｐ１＝｛Ｃ，Ｕ，Φ，Λ｝

4．1．1　概念プリミティブ

対象を観念的に分離・抽象化して得られた認識の単位（カテゴリ）の体系であり、計算機の世界に関する一種の概念分類となっている。体系化は、以下の手順で行なった。

（ｉ）言語資料としたニュース記事１００件に出現した異なり単語約２０００語を分析することにより、要素的な概念カテゴリを抽出する。この際、「共同開発」などの複合語や「使える」などの複合的な概念をもつ語に対しては、複数個のカテゴリを取り出す。ここで、カテゴリ名としては、日本語の単語との区別を明確にするために英単語（省略表現を含む）を用いた。

（ii）抽出された概念カテゴリに対して、類似したものをまとめていき、それら全体に対するカテゴリを設ける。例えば、PERFORMANCE、SPEED、PRICEなどのクラスについてはCRITERION（評価基準）というカテゴリを与える。

（iii）最上位の概念カテゴリとして、SUBSTANCE（実体）、ATTRIBUTE（属性）、RELATION（関係）を考え、これらの下位概念カテゴリを、各カテゴリの抽象度のレベルに応じて決定し、木構造の形式に整理する。

現時点で約７００個のカテゴリを抽出しており、その一部を図５に示す。このように細かく設定した理由は、科学技術文章においてしばしば見られる複合語や格助詞「の」を介在させた名詞句の意味関係の決定を容易にするためである。

図５の中のEVENT-RELATION、COMPOUND-RELATIONとは通常格ラベルと呼ばれているもののクラスであり、原則として、前者は実体概念と属性概念間の、後者は実体概念同士の意味関係を表している。これらのクラスの下位カテゴリの認定については、議論の多いところであるが[10]ここでは文献(11)、(12)を参考にして表１に示すカテゴリを設けた。

４．１．２　判断プリミティブ

概念とは相対的に独立した認識主体の能動的な認識に関するものであり、用例とともに表２に示す。ただし、ここにあげている認識がすべて資料に現われていたというわけではなく、通常の言語活動においてしばしば生じる認識も一部含めている。

４．１．３　認識プリミティブ

認識主体が現実世界をどのように把握したかという認識作用ならびに認識されたひとまとまりの概念（判断）をどのように発展させていくかという思考の展開作用を示したものである。以下に、具体的なプリミティブをその表現例とともに示す。

（１）認識作用［φ１］

（ｉ）事象概念や属性概念を実体化して把握

φ1.11：ある事象に係わる実体の１つを属性をもったものとしてとらえる。

［例］Ａ社が開発したスーパーコンピュータ

φ1.12：ある事象を表１のEVENT-RELATIONに属する概念（時間、場所、目的、原因・理由、様態など）と結び付ける。

［例］処理性能を上げるため
Ｆｏｒｔｒａｎによる場合

φ1.13：ある事象が生起した結果生じた実体と結び付ける。

［例］Ｘを小型化したコンピュータ

表１　概念関係（格ラベル）

関係カテゴリ名		関係の種類			用例
EVENT-RELATION	subj	主体（動作主）			Ａ社が開発する
	obj	対象		対象	ＩＣを作る、価格が安い
	objc	比較対象			従来機より優れる
	objm	相互作用の相手			Ｂ社と提携する
	statef	状態			９型から拡大する
	objf	与え手	変化の起点		Ａ社から供給を受ける
	timf	時間			４月から出荷する
	locf	場所			プリンタから打ち出される
	statet	状態			１００ドルに下がる
	objt	受け手	変化の終点		Ａ社に供給する
	timt	時間			１９８７年まで製造する
	loct	場所			ヨーロッパに輸出する
	result	結果			Ｘを改良したＹ
	tim	時点		時間	１１月に発表した
	timd	期間			１年間遅れる
	loc	地点		場所	Ａ工場で製造する
	locp	通過場所			１０ＭＩＰＳを越える
	purpose	目的・用途			人工知能用に開発する
	ins	道具・手段・方法			ＬＩＳＰで書く
	cause	原因・理由			停電で故障する
	material	原料・材料			半導体で作る
	component	構成要素			ＣＰＵとメモリからなる
	role	役割			上位機種として販売する
	condition	条件			ＣＰ／Ｍで動作する
	form	形式			if-then 形式で書く
	degree	程度			５％下がる
	manner	様態			真向から競合する
	view	見方・観点			性能的に優れている
	content	内容			販売すると発表した
	range	範囲			供与に関する契約
	isa	「ＡはＢだ」のＢを指定			容量が１０ＭＢである
COMPOUND-RELATION	mod	属性を持った実体			１６ビットプロセッサ
	part	全体と部分			コンピュータの主記憶装置
	property	実体の性質			主記憶装置の容量
	value	性質とその値			２５６ＭＢの容量
	rel	関係にある実体			以降（AFTER-rel-TIME）

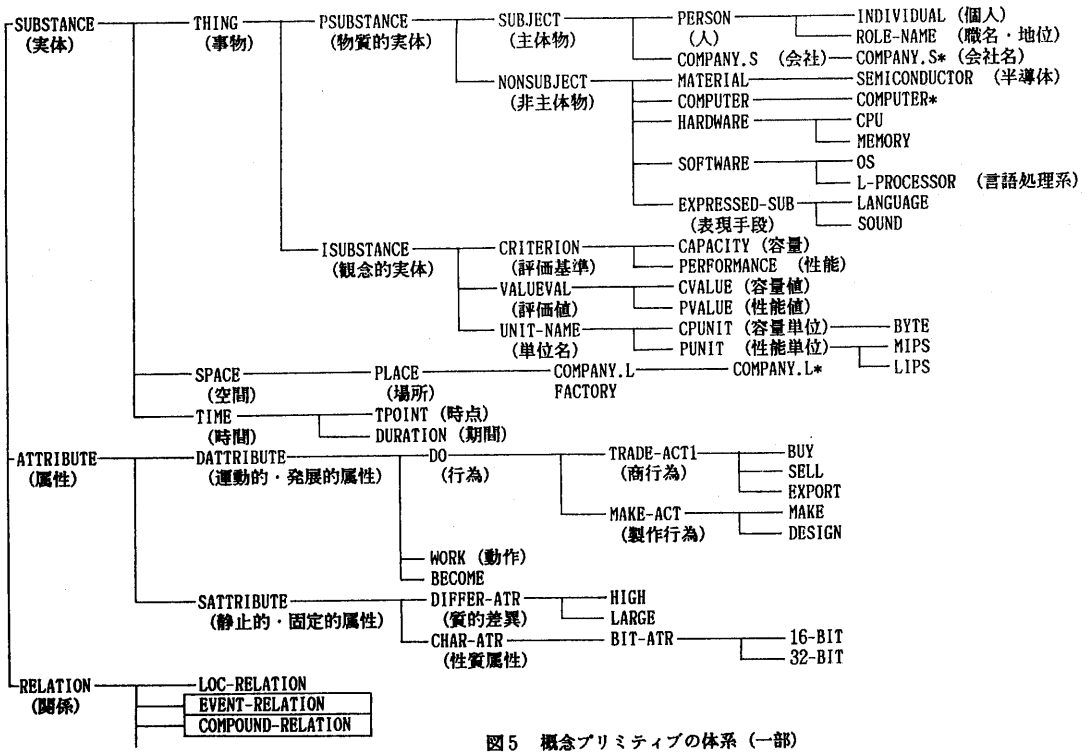

図５　概念プリミティブの体系（一部）

φ1.14：ある具体的な事象概念全体を抽象的に実体としてとらえなおす。

［例］ビジネスで使用するの
ステレオサウンドが聞けること

φ1.15：ある事象概念全体を直接実体的にとらえる。

［例］プログラミングの生産性向上
コンピュータの販売、事務の合理化

φ1.1i(s)：φ1.1i(i=1～4)とφ1.15を同時に作用させてとらえる。

［例］日本語処理機能、提携関係

(ii) 1つの対象をある側面で分離し、多面的に把握

φ1.21：具体的と抽象的とに分離してとらえる

［例］演算処理、ｵﾍﾟﾚｰﾃｨﾝｸﾞ・ｼｽﾃﾑ ＭＶＸ

φ1.22：異なった見方で分離してとらえる。

［例］信号電圧、　走り去る

(iii) 2つ以上の対象を列挙

φ1.31：論理積的に列挙する。

［例］システム／３８とシステム／３６

φ1.32：論理和的に列挙する。

［例］ＦｏｒｔｒａｎまたはＢａｓｉｃ

(iv) 認識ｕｉを対象化して概念として把握

φ1.4 ：ｕ25（仮定）→　仮定する

(v) 実体化φ1.11において属性概念を省略し、直接2つの実体概念を結びつけた形でとらえる［φ1.5］

［例］Ａ社のパソコン、ホームコンピュータ

(2) 思考の展開［φ2］

(i) 2つの事柄を論理的に結合

φ2.11：2つの事柄を順接的に結び付ける。

［例］コンピュータを国内で製造し、4月から出荷を始める。

φ2.12：2つの事柄を逆接的に結び付ける。

［例］ＲＳＸ１１で使用可能であったが、ＶＭＳでも使用可能になった。

(ii) 認識の立場を移行

φ2.21：全体的な把握から細かい部分の把握へ移行する。

［例］・・・コンピュータを発表した。その性能は・・・。

φ2.22：把握するときの見方を別の見方に移行する。

［例］速度は・・・。記憶容量は・・・。

φ2.23：時間的に移行する。

［例］１９８０年には、・・・。ついで１９８５年には、・・・。

φ2.24：空間的に移行する。

［例］日本では・・・。ｱﾒﾘｶでは・・・。

φ2.25：認識の対象を別のものに移す。

［例］E800は・・・。一方E600は・・・。

(iii) 前の事柄に対して、それに付け加わる内容をもつ事柄を結び付ける。［φ2.3］

［例］256MB の主記憶装置が使えるほかに、1GB の拡張記憶装置も使える。

表2　判断プリミティブ

記号	内容	用例
u01	肯定判断	上位機種であるパソコン
u02	否定判断	PL/Iはサポートしていない
u03	普遍判断	Prologは論理型言語である
u04	特殊判断1(提題)	価格は５００万円である
u05	特殊判断2(対比)	A社もパソコンを発表した
u06	当然判断	―――
u11	意志	開発を効率化しよう
u12	推量	工場が使われることになろう
u13	確認	システム用に設計されている
u21	疑問	―――
u22	命令	―――
u23	禁止	―――
u24	勧誘	―――
u25	仮定	もし自社生産するとすれば
u31	観念的な　限定判断	初年度だけで1万台生産する
u32	前提に　例示判断	教育などに使える
u33	対する　超過判断	２．５ＧＢにまで広がった

4．1．4　関係プリミティブ

対象と認識主体との関係を表すもので、時制関係（λ1）と指示関係（λ2）を考えている。

4．2　パターン

前節で述べたプリミティブは、記者の認識構造を記述するためのいわば単位となるものであるが実際の言語活動においては、これらが結びついて複合的な概念あるいは事象として認識されることになる。さらにこれらの概念、事象が結びついてより大きな認識としてまとまっていく。このようなプリミティブ間のまとまりをパターンと呼ぶことにする。パターンの中でこれ以上要素的なパターンに分解できないものを、特に基本パターンと呼ぶ。基本パターンは複合的な概念のまとまりを記述する事物パターンＧと単文で表されるような事象概念のまとまりを記述する事象パターンＶとに大別される。

（1）事物パターン

Ｇ1：2つの実体概念がCOMPOUND-RELATION（一部EVENT-RELATION）で結びついたパターン

（Ｃ1（COMPOUND-RELATION　Ｃ2））

これはＣ1のCOMPOUND-RELATIONがＣ2であることを表しているが、Ｃ1がＣ2のCOMPOUND-RELATIONであるときは、

（Ｃ1（COMPOUND-RELATION(-) Ｃ2）

と書く。パターンＧ1の例を表3に示す。

Ｇ2：認識作用φ1.2で結びついたパターン

（i）（Ｃ1（φ1.21　Ｃ2））

Ｃ1はＣ2の上位概念である。

［例］3081ﾌﾟﾛｾｯｻ　ﾌﾟﾛｸﾞﾗﾐﾝｸﾞ言語Basci

表3　事物パターンG1

RELATION	C1	C2	用例
mod	MA-ATR	COMPUTER	ミニコンピュータ
	BIT-ATR	PROCESSOR	16ﾋﾞｯﾄﾌﾟﾛｾｯｻ
	COL-ATR	DISPLAY-UNIT	カラーCRT
	MD-ATR	MEMORY	256Kﾋﾞｯﾄ RAM
part	COMPUTER	MEMORY	ﾊﾟｿｺﾝ の主記憶装置
	DISPLAY-UNIT	SCREEN	CRTの画面
property	MEMORY	CAPACITY	記憶容量
	COMPUTER	PERFORMANCE	−
	COMPUTER	TYPE	機種
	COMPUTER	PRICE	ミニコンの価格
	COMPUTER	SPEED	−
	PSUBSTANCE	NUMBER	−
value	PRICE	MONEYVALUE	−
	CAPACITY	CVALUE	容量1MB
	NUMBER	COUNTVALUE	−
	TIMEC	TVALUE	ﾏｼﾝ･ｻｲｸﾙ100ns
loc	COMPANY.S	COUNTRY	米国のHP社
material	SEMICONDUCTOR	LASER	半導体レーザ
	MOS	IC	CMOS LSI
purpose	COMPUTER	SOFTWARE	パソコン用OS

(ii) (C1 (φ1.22 C2))

C1、C2が共通の上位概念をもつ。

[例] 文字データ　音声信号

G3：認識作用φ1.3で結びついたパターン（一般に並列句と呼ばれる表現に対応するパターンである）

(i) (φ1.31 C1 C2・・・C3)

[例] RT11とRSX11

(ii) (φ1.32 C1 C2・・・C3)

[例] MS−DOSまたはCP／M−86

(2) 事象パターン

これは、格文法における格フレームに相当するもので、属性概念と共起しうる実体概念あるいは他の属性概念との関係を、概念プリミティブEVENT-RELATION（表1）で記述したものである。なお、属性概念の一部は事物パターンとしても記述されている（表3）。以下に事象パターンの例を示す。

[例] (i) (SUPPLY (subj SUBJECT)
(obj NONSUBJECT)
(objt SUBJECT)
(timf TPVALUE)
(condition OEM))

(ii) (MAGNIFY (subj SUBJECT)
(obj CAPACITY)
(statef CVALUE)
(statet CVALUE))

(iii) (RISE (obj CRITERION)
(degree RATIOV))

(iv) (CHEAP (obj PRICE)
(objc PRICE)
(degree RATIOV))

4．3　プロダクション

基本パターンを変形したり、2つ以上のパターンを組み合せることにより新しいパターンを生成することができる。これはパターンでまとめられた単位的な認識が、多元的、立体的に結びついて1つの大きな認識にまとまる過程に相当する。そこで、これらの変形や接続をプロダクション（生成規則）として整理することにする。

(1) 事象パターンの変形

(i) 事象パターンの具体的な属性を抽象的な属性で補足する。

Rule1.1 (C head (α・・・)・・・)→((C head (φ1.21 Cx))(α・・・)・・・)

ここでαはEVENT-RELATIONに属する概念である。Cheadを補足するCx としては
ABLE（可能）BECOME（変化）CAUSE（使役）
CONTINUE（継続）passive（受動）DO(他動)
TRY（試行）START（開始）などがある。

(ii) 事象パターンで表現される事柄に対して主体の能動的な認識を附加する。

A．事象全体に附加する。

Rule1.2 (C (α・・・)・・・)→(C (MODAL ui) (α・・・)・・・)

ui としてはu01(肯定判断) u02(否定判断) u03(普遍判断) u06(当然判断) u11(意志) u12(推量) u25(過程)などがある。

B．事象パターンの中の特定の実体あるいは属性に対して附加する。

Rule1.3 (C (α・・・)・・・)→(C (α(・・・(MODAL ui)))・・・)

ui としてはu04(提題) u05(対比) u31(限定) u32(例示) u33(超過)などがある。

(2) 事象パターンの事物化

事象パターンでとらえた事象概念を認識φ1.1により事物パターンとして扱えるようにする。

Rule2.1 (Cp・・・(α C1)・・・)→(C1 (φ1.11 α (Cp・・・)))

Rule2.2 (Cp (・・・)・・・)→(Cx (φ1.12 (Cp (・・・)・・・)))

Cx は時間、場所、目的、原因、様態などEVENT-RELATIONに属する概念である。

Rule2.3 (Cp (・・・)・・・)→(Cx (φ1.13 result (Cp (・・・)・・・)))

Cx は事象(Cp ・・・) が生起した結果生じた実体である。

Rule2.4 (Cp・・・)→(φ1.14 (Cp・・・))

Rule2.5 (Cp・・・)→(φ1.15 (Cp・・・))

Rule2.6 上記Rule2.1〜2.4のφ1.1i(i=1〜4)をφ1.1i(s)に置き換えた規則

(3) 事象パターンの接続

(i) 2つのパターンV1、V2を認識φ2により接続する。

Rule3.1 (V1,V2)→(φ2.k V1 V2)
{k=11,12,21-25,31}

(ii) 2つのパターンＶ１、Ｖ２を時間的に同時として接続する。

Rule3.2 (V1,V2)→(# V1 V2)

(iii)パターンＶ１(C1･･･)をRule2.2で事物化しパターンＶ２(C2･･･)の１つのスロットに埋込むことでＶ１、Ｖ２を接続する。

Rule3.3 (V1,V2)→(C2･･･(Cx (φ1.12 (C1 ･･･))))

4．4 記述例

実際のニュース文に対して、メタ言語で記述した認識構造の例を以下に示す。

[ニュース文]
日本電気は１６ビット・パソコン「ＰＣ－９８００シリーズ」の最上位機種として「ＰＣ－９８０１Ｍ３」を発売した。２０Ｍバイトの固定ディスク装置１台と１Ｍバイトのフロッピーディスク装置を内蔵している。ユーザが使える記憶容量は２５６Ｋバイトである。価格は８３万８０００円である。
（日経ｴﾚｸﾄﾛﾆｸｽ 1985 3.11 産業ﾆｭｰｽﾀﾞｲｼﾞｪｽﾄより）

[認識構造]
```
(N0094 (JOUR NEKKEI-E)
       (DATE 1985-3-11)
       (CONT (φ2.21 V1 (φ2.22 V2 (φ2.22 V3 V4))))
(V1 (PRED SELL)                                    [発売]
    (MODAL u13)
    (subj G1)
    (obj PC9800*)                                  [PC-9801M3 ]
    (role G2))
 (G1 (HEAD COMPANY.S*)                             [日本電気]
     (MODAL u04))
 (G2 (HEAD TYPE)
     (property(-) G3))
  (G3 (HEAD COMPUTER)                              [最上位機]
      (mod HIGHEST-RANK)
      (objc G4))
   (G4 (HEAD PC9800)                               [PC-9800ｼﾘｰｽﾞ]
       (φ1.21 G5))
    (G5 (HEAD COMPUTER)
        (mod PERSONAL)
        (mod 16-BIT))
(V2 (PRED (HEAD HAVE)                              [内蔵]
          (φ1.21 CONTINUE))
    (obj (φ1.31 G6 G7)))
 (G6 (HEAD DISK)                                   [固定ﾃﾞｨｽｸ 装置]
     (mod HARD)
     (property G8)
     (property G9))
  (G8 (HEAD CAPACITY)
      (value CVALUE))                              [20Mﾊﾞｲﾄ ]
  (G9 (HEAD NUMBER)
      (value COUNTVALUE))                          [１台]
 (G7 (HEAD FLOPPY-DISK)
     (property G10))
  (G10 (HEAD CAPACITY)
       (value CVALUE))                             [1Mﾊﾞｲﾄ]
(V3 (PRED (isa CVALUE)                             [256Kﾊﾞｲﾄ]
    (MODAL u01)
    (obj G11))
 (G11 (HEAD CAPACITY)                              [容量]
      (MODAL u04)
      (property(-) G12))
  (G12 (HEAD MEMORY)
       (φ1.1 obj V5))
   (V5 (PRED (HEAD USE)                            [使える]
             (φ1.21 ABLE))
       (subj USER))
(V4 (PRED (isa MONEYVALUE))
    (MODAL u01)
    (obj G13))
 (G13 (HEAD PRICE)                                 [価格]
      (MODAL u04))
```

5．言語表現と認識構造との対応

ここでは４．で述べたメタ言語が実際のニュース文においてどのように表現されているかを調べ、ニュース文から認識構造を得るまでの流れを考える。[15][16]

まず、単語辞書には品詞などの情報とともにプリミティブ（複数プリミティブが結合したものも含む）

表４ メタ言語とニュース文との対応関係

規則名	メタ言語		内容
係り受け規則Ｉ	事象パターン		プリミティブで記述された属性と実体の関係、意味的拘束条件に、パターンが文として表現されたときの文法的条件（格助詞など）および関係が必須か任意かの情報を与えたもの(格フレーム)。
係り受け規則II	事物ﾊﾟﾀｰﾝ	Ｇ１	プリミティブで記述された概念間の関係意味的拘束条件に、パターンが名詞句として表現されたときの文法的条件（自立語の品詞、格助詞、体言的接尾語「向け」や「用」の有無など）を与えたもの。
		Ｇ２	２つの語の概念プリミティブの上位・下位関係で判定。
		Ｇ３	接続詞（および、または）、並立助詞（や、と）の有無、ならびに主要語の概念プリミティブで判定。
パターン変形情報抽出規則	Rule1.1		形式動詞（なる、する、いる、ある、できる、みる）、用言的接尾語（れる、られる、される、せる、させる）、無活用形容詞（可能）、動詞（開始、始める）などが述語に接続［係り受け規則の格助詞が変化することあり］。
	Rule1.2		助動詞（だ、ある、ない、よう、う、べし、た）、接続助詞（ば）が接続。
	Rule1.3		係り助詞（は、も）、副助詞（だけ、まで、など、のみ）が係り受け規則Ｉの格助詞に接続したり、置き換わる。
パターン事物化判定規則	Rule2.1		係り受け規則Ｉ、IIの意味的拘束条件を利用。
	Rule2.2		名詞（ため、とき、場合、際、よう）が埋込み述語文節に接続。
	Rule2.3		係り受け規則ＩにresultのＩ項目を付加。
	Rule2.4		形式名詞（の、こと）が接続。
	Rule2.5		係り受け規則Ｉの意味的拘束条件利用。
	Rule2.6		Rule2.1～Rule2.5を適宜利用。
パターン接続情報抽出規則	Rule3.1	φ2.11	連用中止形接続。第２文の必須要素が省略されることが多い。
		φ2.12	接続助詞「が」による接続。
		φ2.21	第２文のobj 要素（または提題要素）が第１文のobj 要素の性質あるいは部分。
		φ2.22	obj 要素が性質や部分を表す概念でありしかも第２文のobj 要素（または提題要素）が第１文のobj 要素と同じ上位概念をもつ。
		φ2.23	提題要素の概念がＴＩＭＥである。
		φ2.24	提題要素の概念がＰＬＡＣＥである。
		φ2.25	第２文のobj 要素（または提題要素）が第１文のobj 要素と同じ上位概念をもつ副詞句「一方」の存在。
		φ2.31	Rule2.2 による埋込みで「ほかに」に接続
	Rule3.2		副詞句「同時に」の有無で判定。
	Rule3.3		Rule2.2 の埋込み語の概念プリミティブで判定。

を与えておく。パターンやプロダクションについては表4に示すようなニュース文との対応関係を抽出した。これらの関係は係り受け規則、抽出規則としてまとめ、構文・意味解析に使用する。

図6が表4の規則を用いてニュース文の解析を行なうときの概略的な流れである。[13][14]形態素解析の部分（自立語辞書約1000語、付属語辞書約130 語）は既に完成しており、現在各規則の整備と残りの部分のプログラム開発を行なっている。

6．むすび

雑誌等の日本語ニュース文を、それが記者のどのような認識構造をへて生成されたかという観点から解析する際の基本的な考え方について述べた。１００件のニュース文の分析より得られた情報を基にして解析プログラムを作成している段階であるが、その特徴をまとめると以下のようになる。

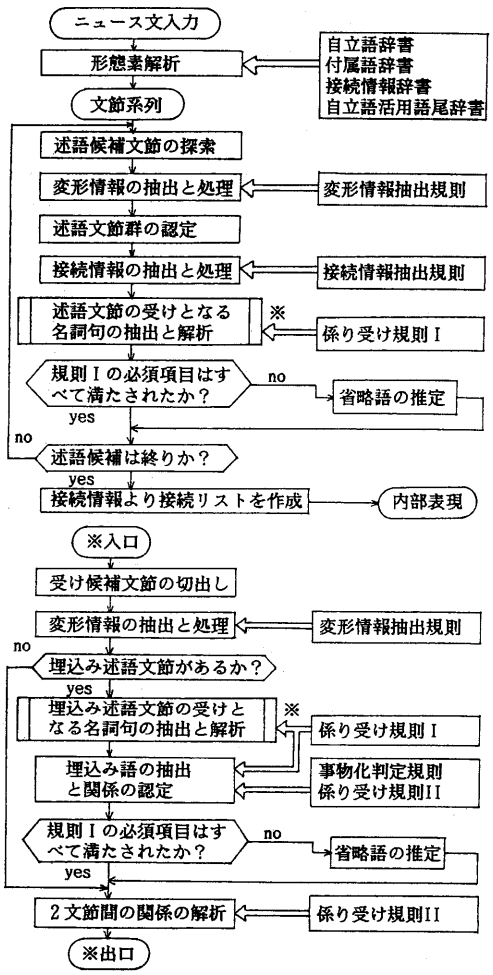

図6　ニュース文解析の流れ

（１）文だけでなく、文章構造の解析まで考慮していること。

（２）表現レベルの構造と認識レベルの構造を分離して解析していること。

図2に示すような情報を抽出するだけならば、記者の認識構造まで考慮しなくても実現可能であると思われるが、本研究では、比較的長いニュース文章の構造を解析し、それを要約することにより知識ベース生成のための情報を取り出すことを最終的な目的としている。そのためには、４．３で述べた事象パターン間の接続関係の記述をより精密化するとともに、それと言語表現との対応を明らかにする必要がある。さらに、「出荷開始は１２月からである。」と「１２月から出荷を始める。」という２文は、今のところ異なった内部表現に対応しており、情報抽出のためには、これらを標準的な形式に変換することも必要になる。現在プログラム開発と併せて、より多くのニュース文の収集と詳細な分析を進めている。

参考文献

1）Hendrix,G.G. et al.:Developing a Natural Language Interface to Complex Data,ACM Trans. Database Syst., Vol.3, No.2, pp105-147(June 1978).

2）Salveter,S.:Supporting Natural Language Database Update by Modeling Real World Actions, Proc.1st International Workshop on Expert database Systems (Oct.1984).

3）藤崎，間下，渋谷，鷹尾：データベース照会システム「ヤチマタ」と名詞句データ模型，情報処理学会論文誌，Vol.20，No.1，pp77-84(1979).

4）佐藤，斉藤，菊地：特許情報検索のための日本語質問文解析，情報処理学会論文誌,Vol.25，No.3(1984).

5）絹川博之：表階層モデルに基づく自然語インターフェース処理方式，情報処理学会論文誌,Vol.27，No.5(1986).

6）絹川，木村：日本語構造解析による自動インデクシング方式，情報処理学会論文誌，Vol.21，No.3(1980).

7）高松，日下，西田：技術抄録文からの関係情報の自動抽出，情報処理学会論文誌，Vol.25，No.2(1984).

8）遠藤勉：日本語によるデータベース照会システム，大分大学工学部研究報告，第９号(1982).

9）遠藤勉：格構造を利用した歴史データの記述と質問応答への応用，情報処理学会第25回全大論文集，3k-9(1982).

１０）辻井，山梨：格とその認定基準，情処学会ＮＬ研究会資料　52-3(1985).

１１）島津，内藤，野村：格構造モデルに基づいた日本語文の分析と解析，情処学会ＮＬ研究会資料　29-1(1982).

１２）坂本義行：格構造を中心とした用言と付属語辞書，情処学会ＮＬ研究会資料　38-8(1983).

１３）遠藤，田町：日本語文章の構造記述，信学技報，AL79-37(1979).

１４）Endo,T.and Tamati,T.:Decomposition of Japanese Sentences into Normal Forms Based on Human Linguistic Process,Proc.COLING80(1980).

１５）遠藤勉：日本語によるデータベースの操作――言語モデル――，電気関係学会九支連大，No.460(1985).

１６）高原，遠藤：日本語によるデータベースの操作――データベースの生成――，電気関係学会九支連大(1985).