自然言語処理 29－1
（1982. 1. 22）

格構造モデルに基づいた日本語文の分析と解析

島津 明　　内藤 昭三　　野村 浩郷
（日本電信電話公社 武蔵野電気通信研究所）

1. はじめに

意味理解の問題を追求するために、知能的機械翻訳システムの実現を具体的な目標として、日本語文の分析および日本語文の解析・意味表現・生成などについて研究を進めている。翻訳システムの対象は科学技術解説文で、日本語については雑誌「サイエンス」の日本人著者による解説文を対象にしている。

自然言語理解の研究が知識表現の研究とともに各所で行われている。従来のこの種の研究は非常に制限された世界を対象にしている場合が多い。しかし、自然言語には、簡単には制限ができないこと、同じことを言うのにも多様な表現が可能なことなどの特質があり、かなり広い範囲で自然言語の種々の面の特徴を捉えることが重要と考えられる。この特徴を捉えるためには、資料の種々の角度からの分析が必要となる。我々はこの分析を言語理解で広く用いられている格構造[1)(3)]に基づく言語モデルを設定し、それに基づいて進めている。また、この言語モデルは解析システムのベースとなるものでもある。我々の格構造モデルの特徴は多くの種類の格を設定していること、単位文間の関係や法情報を重視していることなどにある。以下、2章で言語モデルを、3章で雑誌「サイエンス」の1文献を対象にした種々の分析の概要を、4章で解析手法の概略を述べる。

2. 分析および解析の言語モデル

分析および解析における言語モデルを表1に示す。このモデルに基づき分析が行われ、解析の方略が立てられる。このモデルは事象とその関係を基礎に組み立てられている。個々の事象は単位文を中心に言語表現され、事象間の関係は単位文を連用修飾関係や連体修飾関係で接続することにより言語表現される。1個の事象は用言情報、格関係情報、法情報により表現される。用言情報は行為・状態・関係・属性などの事象の種類を示す。格関係情報は事象の内容を具体的に説明する格要素により表現される。法情報は用言情報や格要素あるいは事象そのものに他の角度からの説明を加える。格要素は名詞句や副詞句とこれに付属する助詞などにより表現される。名詞句・副詞句により、事象に関係する具体的あるいは抽象的な対象、あるいは事象の様態・程度などの情報が表される。法情報は用言に付属する助動詞、助詞、補助用言、あるいは名詞句に付属する助詞により表される。法情報としては、時、相、判断・態度、様態などの情報が表される。

（1）事象　一般に単位文で1個の事象が表され、外側の単位文でその事象を規定・補足する事象や結果として生起する事象などが表される。単位文の出現する順にみれば、最初に場面が設定され、順に場面の説明や場面の時系列的な進

表1. 言語モデル

文章＝文 | 文章＋文章
文＝単位文 | 文＋接続情報＋文
単位文＝格情報＋法情報
格情報＝用言情報＋格関係情報
格関係情報＝対象1，手段・方法などの表2に示す格
法情報＝時＋相＋判断・態度＋様態

展などが説明される。

例）急行電車の窓から駅名表示板を見るとき、簡単な駅名ならばひと目で読みとることができるが、字数が多いとむずかしい。

この例文では先ず「急行電車の窓から駅名表示板を見る」と場面の設定が行われ、後続の単位文で「駅名表示板を読みとることができる」というその場面での行為が述べられている。このような単位文間の関係に対する見方は、単位文レベルに限定されるものではなく、文レベル、文章レベルでも言えることである。ただし、上位のレベルでは情報の整理・統合化がなされる。これはちょうど０１１１１０１１１１０を３桁づつ区切って３６７６と憶えるのに通ずるところがある。

（２）格システム　格システムについてはSchank[3]のようなprimitive actsと一体となって構成される概念的なもの、Celce-Murciaのような５種類の少ない格を設定するもの、Martin[4]のような言語表現の多様性に合せて30種類程度の多くの格を設定するものまで種々提案されている。また、言語学者の格認定法には直観（intuition）[2]が大きな役割を果している。これは格の種類が少ないのも一因ではないかと思われる。

我々の格システムは、Martinのように言語表現の多様性に合わせた多種類の格から成る。これを表２に示す。この格システムは類似した性質や表現形態により類別され、階層性をもつ。例えば、時を表す格と場所を表す格は物理的な場面を設定するという点で似ている。言語理解では、格システムのトップレベルの数種類の格で漠然とした理解が始まり、理解が深まるとともにより下位のレベルの格が認識されていき、また、このような精密な格関係が認識された後、階層的に類別された個々の格の間の関係や格フレーム間の関係により、個々の事象間の関係やエッセンスの理解が進むと考えられる。

（３）事象と格　ところで、実際の文を眺めてみると、同じ内容でも名詞句・副詞句で表されたり、単位文で表されたりする場合がある。

例）・・・ＳＯＡが0.05秒の場合は、知覚の範囲が3.0個にすぎなかったが、・・・

「ＳＯＡが0.05秒の場合」は、「実験のパラメータである刺激提示時間が0.05秒に設定されたとき」を意味する。これと同様の内容が別の文では、「ＳＯＡ0.1秒の場合」と助詞を省略して表現されている。すなわち、動詞であるとか名詞であるとかは意味理解においては本質的なことではないと考えられる。[5]また多くの格システムでは場所や道具・方法の格があるが、これらは単位文中の格要素として表現されることもあれば、外側の単位文あるいは文で表現されることもある。（例えば、「・・・ドット数を横軸にして反応時間をプロットし・・・」）

このような観点から、１個の単位文を中心に見るとき、単位文中の格要素だけでなく、この単位文を修飾する外側の単位文もまた一種の（広義の）格情報を表現していると考えることができる。外側の単位文によって表現される格は、普通、outer case（inner caseに対して）あるいはadvice case（specific caseに対して）などと呼ばれるものである。

３．科学技術解説の日本語文の分析

雑誌「サイエンス」の解説文の分析を進めている。ここでは日本人著者による１解説文（1978.1. 大山 正 "ひと目でいくつのものが見えるか"，実験心理学に関する解説文。以後、

表2．格関係情報の説明

（6個のグループに分け、関連する格の順序に配列）

	格（略号）	簡単な説明と例
グループ1	対象（O）	行為・状態などの対象．「急行電車の窓から駅名表示板を見る…」*
	対象2（O2）	行為などに特有な第2の対象．「問題は…場合だけに限らない」
	形容対象（KO）	形容詞・形容動詞に対する対象．「個人差も多く…」
	比較対象（CO）	状態・属性の比較対象．「ドット数がmより多いとき…」
	陳述対象（DO）	「AはBだ」型の文におけるA．「…対象が人であろうが…」
	陳述対象2（P）	「AはBだ」型の文におけるB．「…人数も5,6名なら…」
	並列対象（HO）	動作主・対象と並列して行為に係る対象．「…筆者は宮本考雄とともに次のような実験を行った」
	説明対象（SO）	対象を説明するもの．「…知覚の範囲は7.3個と通常の値に達した」
	動作主（A）	行為を直接的に起す対象**．「…私たちが、ひと目で把握できる…」
グループ2	手段・方法（I）	手段・方法．「ひと目で把握できる対象の数…」
	原料・材料（M）	対象の形成に用いられるもの．「…電子回路からなる本体…」
	原因・理由（C）	事象の原因・理由．「…その間の忘却で…答えられない」
グループ3	源泉（S）	変化前の状態・位置．「…点のすべてから…選ばれた4～14の位置…」
	源泉2（S）	事象を陳述する出所．「彼らによると…れる過程である」
	方向・目標（D）	方向・着点などの未来の状態．「マイクに答の音声が入る」
	目的（G）	行為の目的．「…の数を知るのに用いられる…」
	結果（R）	対象の変化した状態．「…実験結果はCのようになった」
グループ4	場所・位置（L）	事象が生起する場所．「…の下図に…ローマ字がある…」
	時（T）	事象開始時点・継続時間など．「彼が1871年に…発表した…」
	場合（B）	事象生起の場面．「…SOAが4秒の場合に11個となった」
グループ5	対象区別（TAI）	特に区別される対象．「これに対して、…が用いられた」
	対象関与（TUI）	行為が関与する対象．「第3行について答える」
	対象代替（KAW）	代理の対象．「…意見を聞くかわりに、…調査する…」
	対象資格（TOS）	資格などの対象．「…ことは…の法則として指摘した…」
	対象着目（TOT）	特に着目される対象．「…ドットの知覚にとって…大事な時間…」
グループ6	頻度（RAT）	事象が反復して生起する頻度．「…10試行中5回あやまり、…」
	程度（DEG）	変化や状態の程度．「…実験結果と、…結果がよく一致して…」
	相定（ASU）	事象生起の推定．「おそらく…5字ぐらい当っていただろう」
	形容（MOD）	事象の様態的形容．「…の過程が明瞭に現れたことは…」

＊　例文は分析資料からとった。

＊＊　多くのシステムでは動作主は有生で行為に直接の責任を持つものとされている。「太郎が窓ガラスを割った」という事象では、「太郎」の意志の有無によって「太郎」が動作主格であったり経験者格であったりする。しかしながら、「太郎」は意志に係らず「割る」という行為に関して存在する対象であり、意志の有無は推論されることでもある。このような観点から我々のモデルでは動作主格は対象格の特別のもので、意志の有無は推論の問題と考える。

資料と呼ぶ。）の幾つかの角度からの分析結果を述べる。法情報については、概要を報告した。[6] ここでは紙数の関係で省略する。

文章の理解は全体を解析して初めて完了すると考えられる。その意味で以下に示す量的数値は、言語理解システムがサイエンス程度の文献に対して扱う、あるいは必要とする言語的情報量の目安を与えるものである。

3．1．分析の概要と全体的データ

- **分析の方法**　各文において、用言とその格要素とを認定し、この単位文に係る単位文を認定する。また、各格要素の構文形態を分析する。例を図1に示す。
- **全体的量**　資料は雑誌「サイエンス」の平均的なサイズの文献である。句点で終る文が213個ある。総語数は短い単位で7351、異なり語数は、949である。（「駅名表示板」は、「駅」「名」「表示」「板」と分ける。）
- **1文あたりの単位文数が多い**　単位文は851個あり、1文あたり約4個である。（図2参照）ただし、単独で出現する連用形の用言で、様態などの情報を表すものは単位文として数えない。従来、言語理解で扱われた文では、単位文4個というのは多いほうと思われる。一般の文の複雑さの一端が分る。
- **埋込み文が多い**　従来の言語理解で対象とされた文に比べて多い。（図2参照。1文あたり平均約1.5、多いもので8個）

単位文や埋込み文が多いと、単位文と単位文，単位文と名詞との間の係り受け関係の場合の数が増える。これは名詞句レベルの係り受け関係の組み合せとともに相乗的に増大する。一般の文の複雑さを示す一側面であり、解析上大きな問題となるものである。

- **出現頻度の少ない用言が多い**　用言の出現頻度を図3に、品詞別うち分けを表3に示す。ほとんどの用言（約80%）が高々3回しか出現しない。このことは1文献の処理にも多種類の格フレームなどの言語データが必要であることを意味している。比較的多いのは、「だ」で、これにより2個の対象・概念が結合されて新しい情報が蓄積されていくことが窺える。

3．2．格関係のデータ

- **対象格が多い**　おのおのの格の出現数を表4に示す。対象格の出現回数が多

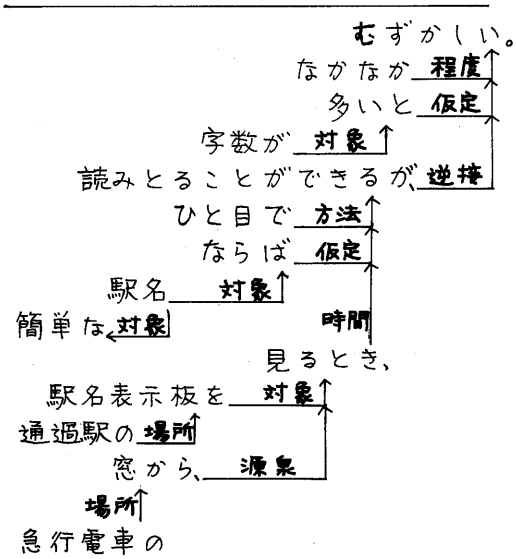

図1．言語モデルに基づいた分析例

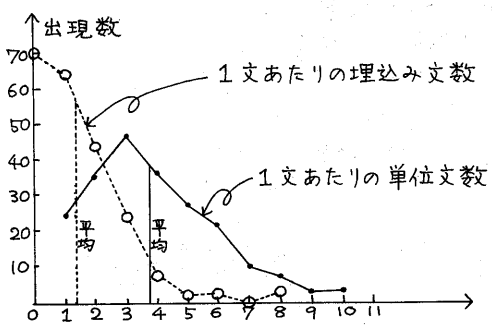

図2．1文あたりの単位文数と埋込み文数

いことは、対象という捉え方に立って新しい情報が加えられて話が展開されていくという日本語表現の一面を反映していると考えられる。

- **単位文あたりの格要素の出現個数が少ない**　平均約1.8個、多くても5個である。このことは、一度述べたことは省略する、共通の格要素は省略するなどの日本語表現の特色を反映している。従来の言語理解システムの対象とした文と対照的である。

翻訳においては、格の省略は大きな問題となる。動作主格が省略され、対象格がある場合は受身文で表現することが可能である。両方とも省略されるか、対象格を本来とらない用言で動作主格などが省略された場合（資料では約8％）が問題で、文脈処理が必要となる。

- **各格に対応する助詞列は多様である**　例えば、対象格に対応する助詞列には、「を」「でも」「しか」「が」「も」などがある。全部の格に対しては82種類の助詞列があった。解析のためにこれらの言語データを蓄積・整理することが重要である。

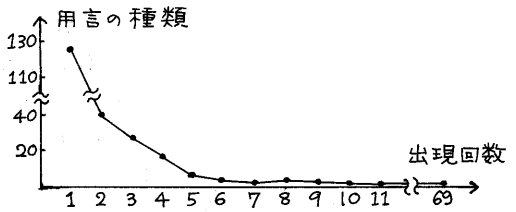

図3．用言の種類と出現回数

表3．用言の品詞別の出現数

品　詞	異なり語数	総語数
動詞	213	657
形容詞	20	56
形容動詞	21	41
だ（だ,である）	1	90
体言止め	1	7
計	256	851

- **「を」は対象格だけを表す**　一般に各助詞列は複数の格に対応するが、「を」は対象格だけに対応している。このことは格の解析に寄与するものと考えられる。
- **1個の用言の同じ格に対しても種々の名詞句・副詞句が表れる**　例えば、「用いる」の対象格としては、「黒点を」「スライドが」「黒豆を」「それを」「ドット刺激提示装置を」「反応時間を」「過程」（被連体修飾語）「視覚イメージを」など多様である。画一的な意味カテゴリの表現と単純な適用では不十分であり、言語データ・知識ベースの蓄積と柔軟なマッチングが重要である。

3．3．名詞句・副詞句（格要素）のデータ

- **全体的量**　用言に左から係る名詞句（付属する後置詞は除く）・副詞句は、1310個（異なり数958）である。副詞句あるいはそれに相当するもの（「明らかだ」の連用形「明らかに」など）は様態情報を表す格（表2の第6グループ）として扱っている。
- **名詞句の構文形態の種類が多い**　（Nの（NのN））、（－　P　N）などの名詞句の構文形態の種類が多く（98種）言語データの充実が重要である。（表5参照）ここでN、P、－はそれぞれ名詞、用言、格要素の省略を表す。例えば、「アパートホームにいる人」は、（－　P　N）となる。

しかし、頻度10位までの構文形態でほとんど（89％）を占めている。

表4．格の出現回数　（表2の第5グループはK、第6グループはY、O,O2,KO,CO,HO,SOはO、S,S2はSとしてまとめて示す。）

格名	O	A	DO	P	I	M	C	S
出現数	619	201	57	97	62	0	4	16

D	G	R	L	T	B	K	Y	計
12	4	44	93	89	40	13	187	1538

- **プリミティブな構文形態も多い**　他の名詞句を生成できる規則となるプリミティブな形（NPのNP，NPとNPなど）が比較的多い（44個）。これは自然言語の個別性という特徴を反映していると考えられる。
- **解析上問題となる埋込み文を含んだ名詞句**　言語理解上重要なこの形態の出現頻度は少なくない（名詞句全体の22%）。埋込みにはいくつかの形態がある[7]。これを表6に示す。
- **係り受け関係の解析が問題となる形態**　・「の」を2個以上含む句(4.5%)　例）無数の文字の断片　・指示語や文脈参照語（「前述」など）と「の」を含む句（2%）　例）その8個以下のドット　・埋込み文と「の」を含む句（3.9%）　例）…役立っている視覚イメージの存続時間　・埋込み文と複合名詞からなる句（2件）　例）英国紳士が黒豆を用いて行った実験結果（「行った」は「実験」を修飾）

3．4．「の」で結合される名詞の連関

- **「の」を含んだ名詞句が多い**　名詞句は名詞あるいは複合名詞だけから成る場合が最も多く（52%），次が「の」を含んだ場合（43%）である。

表5．名詞句の構文構造と頻度（上位10個）

1．N	…607	2．－PN	…186
3．NのN	…182	4．NN	… 78
5．PN	… 33	6．－P(NのN)	…22
7．(NのN)のN	…19	8．(NN)のN	…15
9．(－PN)のN	…14	10．Nの(NのN)	…13

表6．埋込み文の形態別頻度

被連体修飾名詞が埋込み文の格要素…56%
被連体修飾名詞が埋込み文の格要素を修飾…3%
その他の場合（「…すること」など）…41%

「の」で結ばれる名詞句の分析は、セマンティックネットワークなどの知識ベースの構成とも関連しており重要である。

「の」で結合された名詞を調べるにあたり、「の」を含む句で入れ子になっているもの（約14%）は簡約化を行う。例えば、((24ページの上図)の黒線)は(24ページの上図)と(上図の黒線)とする。ただし、この簡約化で特殊なもの（約2%）は除く。例えば、((1個から2個まで)の対象)のような部分句全体で1個の名詞を修飾しており、部分句を1個の名詞に簡約しがたいものである。以上のようにして、(N1＋の＋N2)なる形のもの327個（N1, N2の異なり語は326）が得られる。

例）N1＝ドットパターン
　　N2＝情報、VIS、提示時間、提示、例

- **「の」で連結される名詞のネットワーク**　「の」で連結される名詞・複合名詞はネットワークを形成する。全部で40個の孤立した部分ネットワークができる。その内のひとつだけが大きく（221語よりなる。図4参照）、他は小規模（2～7語）である。最大のネットワークにはループが含まれている。

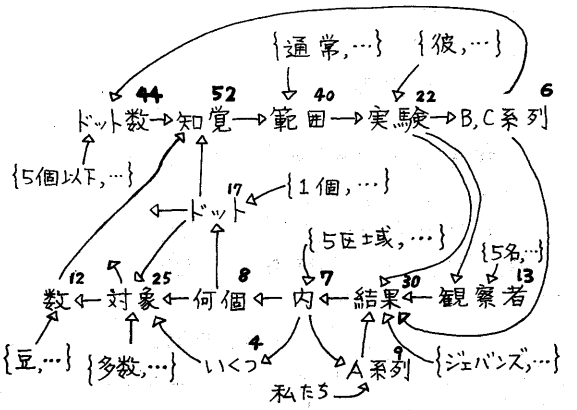

図4．名詞のループを含むネットワーク（最大のもの）＊右肩の数字は出現頻度

ルー7°上にある語は、「知覚」「範囲」「実験」「B,C系列」「観察者」「結果」「内」「A系列」「何個」「いくつ」「対象」「ドット」「数」の13語であり、どれも話題に強く依存した語である。

ルー7°上にある語、枝の出入りの多い語、出現頻度の多い語などと、それらの語が関係する用言を基に、文献の主題の抽出、抄録の作成、知識の習得などが行えると考えられる。

4. 解析手法の概要

日本語文の解析は2章で述べたモデルに従う。また、3章で述べた科学技術解説文の性質や問題点が解析の個々の点を規定し特徴づける。

(1) 考慮すべき言語現象

- 1文あたりの単位文数、埋込み文数が多いので、どの単位文がどの単位文や名詞句に係るか、どの名詞句がどの用言に係るか(格要素の飛びこし)が問題になる。特に、(用言連用修飾句+…+連体形用言+名詞句+…+用言+…)などの文では、下線の用言連用修飾句が右方のどの用言に係るかにより、言語構造の異なる英語などに翻訳した結果も異なってくる。
- 単位文あたりの格要素の出現数が少なく、対象や動作主の格要素が省略される場合が多い。1個の文の解析にも文脈情報や知識の利用が重要である。
- 自然言語を特徴づけ、また翻訳の正確さや質に影響を与える法情報の出現頻度が多い。
- 複合語が多い。複合名詞は本質的には助詞などが省略された名詞句の特別なものである。意味理解の観点からは場合によりそのような解釈をする必要がある。また、埋込み文が複合名詞の一部に係る場合があり、解析上注意を要する。

複合用言も多い(「群化しやすい」など)。複合用言の場合、一般に格フレームの内容は用言が単独で現れる場合と異なる。例えば、「群化しやすい」の場合、形容詞的性質が加わる。

- 名詞句の構文形態は多様である。「の」を複数個含むもの、指示詞などを含むもの、並列句を含むもの、埋込み文を含むものなど解析上注意を要する因子が多い。

(2) 解析の基本枠組

上述のような文の特徴を考慮し、解析の基本枠組は、・辞書や知識ベースに依存した処理をすること、・話の場や事象間の関係・流れを取り込んだ処理をすることにある。

話の場は、省略や指示語の処理、語の意味や用法に関する処理に係わると同時に、局所的な文情報や知識の影響を受けて形成される。格フレームを含む辞書項目、知識(一般常識、専門知識)、解析の状態や結果を表す内部表現などはフレーム的記述を基本とする。また、入力文の内部表現には、入力文の意味内容を表現するとともに、表層の形態(スロットの省略の有無、助詞などの構文形態)の痕跡も残す。

(3) 解析手順の概要

解析は基本的に横型で、各段で複数の可能性を残しながら次の段に進んでいく。各段で完結しない処理は後段で補完する。ここで言う横型は単なる部分の積み上げという意味ではなく、意味理解は全体と部分との相互作用によって進んでいくという考えを追求する。これは、単位文数の多さ、格要素の飛びこし、省略・指示語の処理を考慮することによる。このような観点から、解析は文章レベルで捉える。

文章レベルでは、文レベルの処理で得られた場や事象関係の情報をまとめ、それまでに得られている情報との統合、や知識との同化、文レベルの解析の補

完などが行われる。

文レベルでは形態素・句・単位文・文の順で格フレームなどに基づいて解析が進められる。以下に文レベルの各段の処理の概略・特徴・留意点などについて述べる。

i) 形態素レベル　辞書に基づいて非分ち書き文の語（慣用句を含む）の分割、語連接の判定、複合語の処理を行う。

ii) 句レベル　用言部と用言部との間の名詞句・副詞句を構文パタンにより検出し、知識ベースに基づいて意味処理を行う。(埋込み文＋名詞句)の形の名詞句における用言と名詞との関係は単位文レベルで処理する。法情報に関しては、句に付属する副助詞で表される様態情報を抽出する。

iii) 単位文レベル　各用言のもつ格フレームと格要素（名詞句＋付属語列、副詞句）とを照合する。また、述部の主用言を除いた助詞、助動詞、補助用言などにより表現される相、時、判断・態度の法情報を解析する。

格フレームには用言に固有の格のリスト（一般に複数個）が、構文情報や意味情報とともに与えられる。

格の照合は用言の左に向って進められる。照合操作は、格要素の格助詞や副助詞などがもつ格情報による格スロットの選択、意味情報や構文情報の参照などにより行われる。一般に知識や入力文が完全ではないことを前提に排他的な照合は行わない。

3章で述べたように格要素の省略は頻繁に起る。この省略に対しては、ローカルな場面フレーム（Local Scene Frame）を設定して、このフレーム上で格フレームと格要素との照合を試みる。LSFは、場所、時、動作主、対象などの要素のインスタンスを保持するもので、解析の進行に伴って形成される。LSFは場合によっては文を越えて利用され、場や事象の流れを記述するフレームに昇華されていく。

以上の解析に基き格フレームのインスタンススロットが埋められ、法情報とともに単位文に対する内部表現が作られる。

iv) 文レベル　単位文と単位文との間、時・場面などを表す格要素と単位文との間の接続関係を解析する。この解析では、構文形態、事象間の関係に関する知識、格関係の3面から解析を進める。構文形態は読点情報を含めた係り受け的情報である。事象間の関係に関する知識は、（仮定条件 → 行為可能）などの常識的知識等である。格関係は、2章で述べたように、係りの単位文あるいは格要素と、受けの単位文との間の格関係である。以上の解析に基き文に対する内部表現が作られる。

5. おわりに

自然言語を捉えるには広い範囲でみる必要があるという観点に立って、格構造に基づいた言語モデルを設定し、科学技術解説文の分析、ならびに辞書を含めた知識ベースを核とする解析システムの作成を進めている。本報告では言語モデル・分析・解析の概要を述べた。現時点で、名詞句の構文形態などの言語的諸情報とその1文献あたりのおよその量とが次第に明らかになってきている。現在、格フレーム、名詞句の構文形態、付属語の形態などの言語データおよび理解に係わる知識構造をより明確化し、蓄積するために、分析内容を深め、資料を増やしている。

文献 (1) Bruce, B.C, "Case Systems for Natural Language" AI, vol.6, pp327-360 (2) Cook, W.A, "Case Grammar: Development of the Matrix Model (1970-1979)", 1979 (3) Fillmore, C.J, (訳) 田中他, "格文法の原理", 1975 (4) Martin, W.A, "OWL Notes", 1974 (5) Minsky, M, "知識を表現するための枠組" in "コンピュータビジョンの心理" (ed.) Winston, P.H, (訳) 白井他, 1979 (6) 内藤, 島津, 野村, "日本語文における法情報の解析", 自然言語処理 26-2, 1981 (7) 奥津, "生成日本文法論" 1974 (8) Schank, R.C, "Conceptual Information Processing." 1975