

格構造を中心とした用言と付属語辞書

坂本義行
(電子技術総合研究所)

概要

本研究は、機械翻訳システム¹⁾における、日本語の形態素解析²⁾、構文解析³⁾に必要な辞書システム⁴⁾の開発を目的としたものである。翻訳の対象としている科学技術文献から標本として JICST の文献抄録電気編 1,000 文を用い、その出現単語について詳細な分析を行い、動詞、形容詞、形容動詞辞書の記述形式を決定した。特に翻訳における構文解析の中心となる格フレームを設定した。実例から各動詞とろの表層格、それが平下深層格および名詞意味コードとの関係も調査し、用言として、動詞、形容詞、形容動詞、付属語として助詞、助動詞の台帳を作成した。翻訳のための辞書引取りソフトウェアの基本部分の設計と開発をすすめている。

1. はじめに

辞書を大きく2つに分け、形態素解析に必要な情報と構文解析等に必要の情報として独立に辞書項目を決定した。従って、形態素詞と構文品詞とで重なった部分がある。

辞書データベースを作成するための作業フォーマット(表1)とその作業用マニュアルを作成した。

特に格フレーム付け作業は、サンプル文の KWIC から後述するいくつかの手順をへて記述する方法を採っている。

計算機への入力形式は、LISP の S 式と決め、その構造で直接パッチが行えるようなパッチフォーマットを作成し、作業フォーマットからの転記する方法を採っている。

名詞の意味コード付けは、名詞辞書⁵⁾

と KWIC を参照して作業をすすめている。

辞書引取りソフトは、現在 UTI-LISP 漢字版を用いて作成されており、LISP マニュアル(Symbolics)への移行も予定されている。

2. 辞書台帳の記述形式

2.1 見出し情報

全ての品詞に共通の記述形式とする。

(1) 見出し語 表記は、漢字かひらがなとする。活用語については、終止形で表わす。複合語も独立の見出しとし、子見出しは立てない。

(2) 語尾字数 終止形における語尾の字数とする。語幹がない用言と特殊活用の助動詞の場合には、見出し語に等しいものとなる。

本研究は国の科学技術振興調整費による「日英科学技術文献の速報システムに関する研究」の一部として行ったものである。

研究協力者として、木村睦子氏(計量計画研究所)、横田匡彦氏(東洋情報システム)、松本氏に、本研究の委託を受けて、その遂行のために、自然言語処理の専内家からなる辞書作業分科会、さらに辞書ワーキンググループ(JICST)と機械翻訳プロジェクトワーキンググループ(京大)を組織して、その審議、指導のもとに本研究をすすめている。

(3) 語基 複合語における構成要素の位置を示す。

(4) 異形語 意味が同じで、読みも同じで、形が異なるものを記述する。送り仮名の異なるものも含める。

(5) 派生語 見出しから派生した、品詞の異なるもの、ただし、形容詞において、「語幹+ナ」のように、ほとんどの見出しにつくものは除外する。

(6) 関連語 類義語、対義語、その他として、「増える」に対して「増やす」、「新しい」に対して「真新しい」等ものを。「見出し」の後に類別を入れる。

2.2 形態素情報

全品詞共通の形とし、形態素処理に関して必要な品詞、活用等の情報を選択方式で記述する。これについては、文献2で説明してあるので省略する。

以後の各項目は、品詞によって、その記述形式が異なる。

2.3 動詞の構文・意味情報

分野別、格パターン別に分けて、分野コード、構文品詞、格パターン、変換見出し、アスペクト、態、意志、意味コード、シソーラスコード等について記述する。

(1) 分野コード 基本的な動詞、分野別の動詞の区別を行う。現在は、出典先を示すにとどめる。従って「電気」としている。

(2) 構文品詞 構文解析上必要な品詞又は細分類を示す。

(3) 格パターン 格変態関係を分類し、英語における Hornby の「格パターン」と同様の分類コードを付与する。現在は、個々のコードでなく、動詞別に一連番号をつけた。

(4) 変換見出し 構文変換に用いられる辞書⁶⁾の見出しへの標識をつける。

(5) アスペクト 瞬時、継続、状態のいずれかを判別し、0で囲む。同一パターン内では、1個しかもつ得ない

表1 動詞辞書作業フォーマット(7/1)

見出し語	品詞	格	項目	電気							
				表型格	深型格	名詞	必				
見出し情報	語尾字数	構文品詞	格	動詞							
	漢字部			V 1		V 2					
	語尾			変換見出し							
	読み			アスペクト							
	異形語	派生語		関連語	態		瞬時/継続/状態		瞬時/継続/状態		
					意志		受身/可能/使役/'である'		受身/可能/使役/'である'		
					意味コード		有/無		有/無		
					シソーラスコード						
	備考			備考							
	形態素情報	品詞		格	項目	表型格	深型格	名詞	必	表型格	深型格
意味コード			意味コード			意味コード	意味コード				
動詞活用型		五	上	一		下	一	サ変	ザ変	カ変	力変
		ア	カ	ガ		サ	ザ	タ	ダ		
		ナ	バ	マ		ラ	ワ				
		グナ	グナノ								
動詞活用型		形	形	動		動	特				
		格	格	接		副	並	終	準		
語尾活用型		体	動	形		形	動				
		前接									
備考			備考								

ものとする。

(1) 態 受身, 可能, 使役, 「である」の内, その動詞に生じうるもので, しかも格パターンに差換をもたらしうるもの全てをOで囲む。

なお, 主体が有意味体の場合, 一般的には「使役」が生じうるわけであるが, テキストの性質上そのような例はほとんどないものとみて, 付けない。

2.4 格支配情報

格支配付け作業マニュアル(図1)

によって得られた情報で記述する。

(1) 表層格 格助詞および格助詞相当語として,

が, を, に, と, で, から, より, へ, まで, として, において, によって, のために, 并(裨格)を記入する。

ただし, 深層格が同一のものは, 同じ表層格の欄に並べて記述する。

(2) 深層格 格支配付け作業マニュアルを用いて, 実際にサンプル文のKWICから, 各動詞の表層格の格ラベル(深層格)を決定し, 格ラベル表(表2)を参照して, その略称(英3文字)を記述する。

(3) 名詞の意味コード テキスト中の動詞および格助詞と結合して, 出現した名詞について, 名詞辞書を引く, その意味コードすべてを用例シートに写し, その用例をもとめて名詞意味コード欄に記述する。

(4) 必須性 格が必須格か自由格かを判別し, 必須格の場合には「1」を, 自由格の場合には「0」を入れる。なお, 必須であるか, 必須でないかの基準は, 現実に出現する率が高いか低いかに依らず, その動詞が潜在的にでも持つであろう格を必須格とする。

1. 目的: 個々の動詞・形容詞・形容動詞について, その動詞等とりうる格助詞(表層格)と格ラベル(深層格)とを記述し, かつそれらの格助詞の前に立ちうる名詞の意味分類をも指示する。

例: 保つ 装置が 水そうを (所定の)水位に
(が/主体/1.4) (を/対象/1.4) (に/状態/1.1)

2. 作業手順

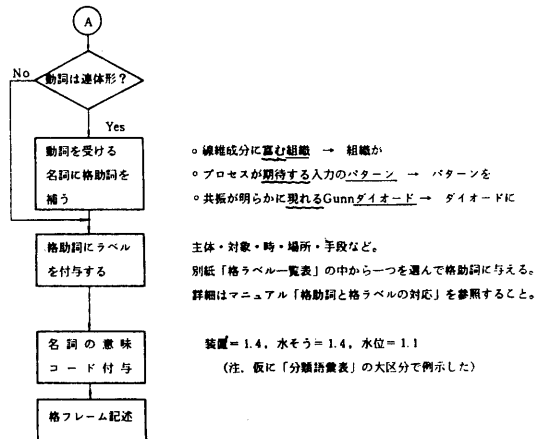
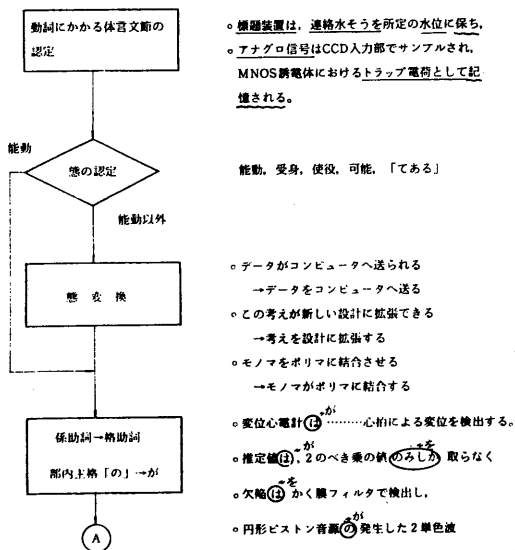


図1 動詞KWICデータから格フレームを記述する法

表 2 日本語格ラベル一覧表

日本語名	英語名	用 例
(1) 主体	SUBject	～が
(2) 対象	OBject	～を
(3) 受け手	RECIpient	～に与える
(4) 与え手	ORIGIN	～から受ける, 奪う
(5) 相手 1	PARtner	～と協議する, 異なる, ～に関連する
(6) 相手 2	OPPonent	～から保護する, 独立する
(7) 時	TIME	1980年に
(8) 時・始点	Time-FRom	5月から
(9) 時・終点	Time-TO	来年まで
(10) 時間	DURation	5分間加熱する
(11) 場所	SPAcE	～に位置する, ～で発生する
(12) 場所・始点	Space-FRom	～から帰る
(13) 場所・終点	Space-TO	～へ送る, ～に到達する
(14) 場所・経過	Space-THrough	～を通る, 上空を飛ぶ
(15) 始状態	SOUrce	} 5.5%から6%へ引き上げる } 英語から日本語に翻訳する
(16) 終状態	GOAl	
(17) 属性	ATTribute	適応性に富む, 欠ける, 之しい
(18) 原因・理由	CAUse	事故で死ぬ, ～から分かる
(19) 手段・道具	TOOl	イオン法で, ドリルで
(20) 材料	MATerial	ペーストで作る
(21) 構成要素	COMponent	～から成る, ～で構成する
(22) 方式	MANner	並列に, 10 m/secで
(23) 条件	CONdition	焦点深度で決まる
(24) 目的	PURpose	～に適する, 備える, 必要な
(25) 役割	ROLe	議長に選ぶ, ～として用いる
(26) 内容規定	COntent	～と呼ぶ, 述べる, みなす
(27) 範囲規定	RANge	～について, ～に関して
(28) 提題	TOPic	～は, ～とは
(29) 観点	VIEWpoint	立場から, ～の点で
(30) 比較	COmparison	～より大きい, ～に劣る, ～を上回る
(31) 随伴	ACCompanyment	～とともに, ～に伴って
(32) 度合	DEGREE	5%増加する, 3キロやせる
(33) 陳述	PREdicative	～である

注. 英語名中, 大文字の部分(3字)を略称とする。

2.5 共起情報

特に動詞と強く結合して出現する慣用的表現等を記述する。

副詞 例 よくいく, よく一致する, よくで"きる。

連用句 例 火を見より明らかだ。

慣用(名詞)
例 手に入れた, 焦点をあわせた, 一致をみる。

2.6 形容詞, 形容動詞の構文・意味情報

(1) 品詞細分類 情意, 性質, 状態に関する情報のいふれかえりで"囲む"。

・情意: 「精神+がる」で動詞化するもの

例 うれしい, 悲しい, 残念だ。

・性質: 「もの」の性質を形容する傾向の強いもの

例 固い, 軟い, 新しい, 細い, きれい, 四角い, 失礼だ。

表3 形容詞、形容動詞辞書作業フォーマット

構文・意味情報	分野コード	形容動詞	
	構文品詞	形容詞	形容動詞
品詞細分類	情態/性質・状態/関係	情態/性質・状態/関係	
変換見出し			
格パターン			
程度性	尺度/極	尺度/極	
名詞性程度	++/+/--/--	++/+/--/--	
意味コード			
シソーラスコード			
備考			

・状態: 「屬性」や「動作」を形容する傾向の強いもの

例 著しい, 甚だしい, 早い, 強い, 十分だ, 困難だ, 正確だ。

・関係: 「ニヒ」, 「もの」の関係を示すもの

例 遠い, 近い, 等しい, 親しい, 同じだ, 異質だ, 反対だ。

(2)程度性 尺度性又は極性のあるものは○で囲む。

・尺度性: 「非常に」, 「かなり」と共起可能な場合。

例 勇い, 大きい, 新しい, 厚い, 活発だ, 一般的だ, 大幅だ, 大だ。

・極性: 「全く」, 「ほとんど」と共起可能な場合で比較級, 最上級がない場合。

例 等しい, 正しい, ない,

同じだ, 一定だ, 一樣だ, 最大だ

(3)名詞性の程度

形容動詞について, その名詞性の強さを4段階として,

・++ ; 名詞性が極めて強く, 「語幹+に」の形で出てきたとき, 概ね(動詞「ある」「なる」に係る場合を除く。以下同様), 「名詞+格助詞」の解釈を受けける。

例 困難に, 健康に

・+ ; 名詞性が強く, 「語幹+に」の形で出てきたとき, 「名詞+格助詞」の解釈を受けけることが多い。

例 毎日に, 幸福に

・- ; 形容動詞性が強く, 「語幹+に」の形が出てきたとき, 形容動詞連

用形の解釈を受けることが多い。
 ・-- ; 名詞性がなく, 「語幹+に」の形が出てきたとき, 必ず形容動詞連用形の解釈を受ける。
 例 ...的に, 静かに, まれに。

3. 翻訳用辞書引込システム

翻訳用辞書のベースとなる辞書構造と辞書引込ソフトウェアの基本設計を述べた。

以下, 辞書引込システムのプロトタイプとして, サンプル文に対する辞書データベースのつくりを述べる。

辞書引込システムの全体図を図2に示す。

3.1 辞書引込システムの概要

辞書引込システムは, 翻訳システムからの要請によって, 指定された見出しに対して辞書を探索し, 所要の情報を引込渡す機能を有している。

辞書引込システムは, この機能を満足するため, ソフトウェア及びソフトウェアにより参照されるデータより構成される。ここでは, それぞれ, 辞書ソフトウェア, 辞書DBと呼ぶこととする。システムの概念図を図3に示す。

3.2 辞書引込ソフトの機能

辞書引込ソフトは大きく分けて4つのフェーズを有する。

(1)「インイテライズ」 システム起動時において, 辞書DBよりDBのコントロール, 探索に必要なDATA, 辞書データを読込込みベクトル化して主記憶上に保存する。

(2)見出し引込 翻訳システムからの, 辞書引込の要求が来たとき, 「語」の同定のため呼び出される。辞書の見出しに存在するか否かの返答を与える。

(3)辞書属性データの引込渡し 見出し引込で, 辞書に存在することが判明している見出しについて, その属性データを翻訳システムに引込渡す。

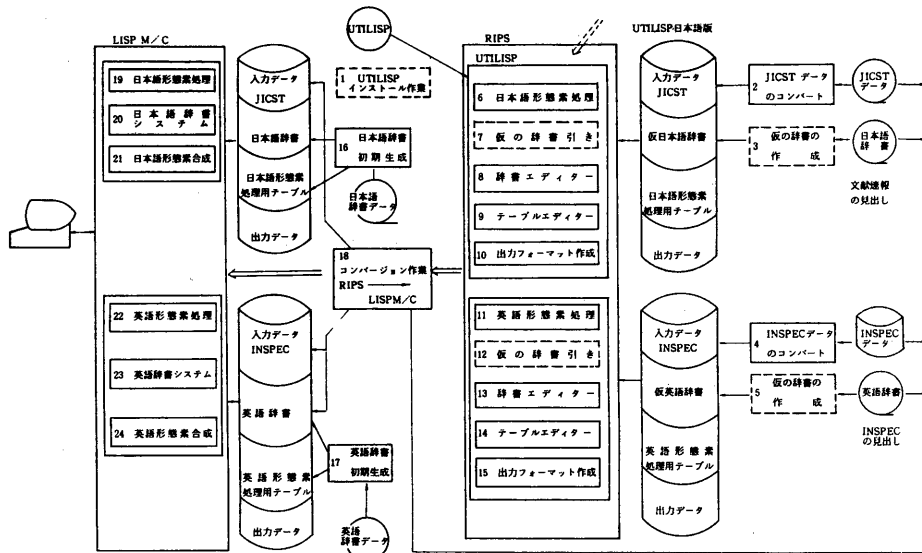


図 2 翻訳辞書システム構成図

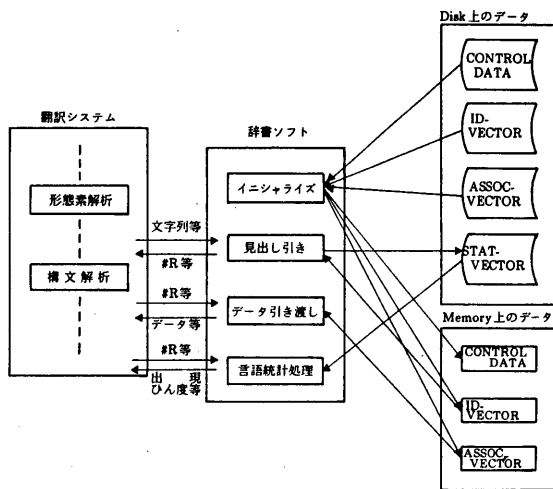


図 3 システム概念図

(4) 言語統計処理 翻訳システムの維持、管理、Version up 等に用いるため、翻訳システムから呼ばれた内容、回数、使用された辞書の内容等についての統計処理を行う。

4. おわりに

本研究は、57年度に基本設計を終え、58年度から開発に着手した段階である。形態素解析、構文解析のテストを繰り返

した後、改良、拡張を方々行っている。

辞書システムについては、データの拡張による大規模化を考慮して、詳細な設計を行っている。

それらについては、漸次、本会報にて報告する予定である。

参考文献

- 1) 長尾真, 「科学技術庁機械翻訳プロジェクトの概要」.
 - 2) 坂本義行, 「日本語形態素解析の基本設計」.
 - 3) 辻井潤一, 「日本語構文解析」.
 - 4) 中村順一, 「文法記述用ソフトウェア GRADE」.
 - 5) 中井浩, 「語の収集と体言を中心とした辞書について」.
 - 6) 西田覚明, 「変換過程の基本設計」.
 - 7) 矢田光治, 長尾真, 「機械翻訳総合システムの基本設計」.
- 以上の各文献は、本会で同時に告されたものである。