自然言語処理　46－1
（1984. 12. 7）

機械翻訳用辞書構成

村木　一至
（日本電気㈱ C&Cシステム研究所）

1. はじめに

実用的な言語処理システムには，大規模語彙辞書が必要である．そのためには，辞書内容を均質に且つ矛盾なく開発・管理するための方策が必要となる．また，機械翻訳の如く，複数言語を対象とするには，複数言語間の語彙に関する対応の管理も重要な課題となる．

本稿では，日英機械翻訳システムVENUSの語彙辞書構成〈三層辞書構成―形態・構文・意味辞書〉とその管理について報告する．[2]

一般に辞書と言うとき，様々な内容の，様々な目的をもつ辞書を含む．言葉の持つ内容，言葉の言語的性質，特定語に関連する言葉について書かれた多種類の市販辞書がある．

言語処理用機械辞書にしても，ワープロ用カナ漢変換辞書，言語解析用文法情報辞書，情報検索用類語辞書など多種類存在する．また，カナ漢変換用辞書でも，その対象によって人名辞書，地名辞書などが存在する．これら様々な辞書は，その目的とそれを直接扱う機械（通常ソフトウエア）の違いにより，内容，形式が異なる．

機械翻訳用辞書は，原言語解析・目的言語生成辞書と，原言語と目的言語対応付け辞書（ないし機能）を全て含んでいる．この辞書内容は，翻訳方式や言語解析・生成方式によって異なる．またその構成も処理方式によって大きな違いを見せる．

言語処理や翻訳方式（システム）に依らない大規模辞書を開発しようという研究が報告されている．[1] そこでの辞書開発方針は，共用利用者辞書と，目的機械辞書をトランスレータにより対応付けようとする．トランスレータを多種類用意することで，同一言語に関する解析・生成用辞書を共用辞書から自動生成する．たとえ翻訳方式の差違があっても，各翻訳システム用にトランスレータを用意して対処する．

機械翻訳の言語対応付けに於けるイディオム記述の問題を解決しようという報告がある．[3] そこでは，単語辞書の1エントリー内が，見出し語に関連する派生語，イディオム，形態情報，構文情報，意味情報を記述するためにどのように構造化すべきかを述べている．

ここで提案する翻訳用辞書構成は，更に以下の問題に対する解決を目的とする．

1) 同義語記述
2) 多言語間対応記述
3) 記述重複の解消
4) 多様な応用への適合性

その他，5)イディオム記述とその処理の簡素化，6)解析・生成という処理方向からの独立などを可能とする．

更に，一旦開発された辞書構成に対応する辞書管理システム（辞書エディタ）は上記6項目の要請を満足させる辞書を保守・管理するために必須である．

以下の節では，本稿提案による辞書構成が生まれた背景を説明し，辞書構成，管理システムについて言及する．

2. 機械翻訳用辞書構成の要件

本節では，筆者等が開発している機械翻訳システムの説明を通し，辞書構成に関し何が要求されるのかを明らか

Kする．

2.1 VENUS機械翻訳システム

VENUS翻訳システムは，入力日本語文を形態素解析し，単語同定を行う．この処理では，熟語判定や未定義語処理を含んでいる．形態素解析は，単語列を生成する．

意味抽出は，入力単語列から文法情報と意味情報を用いて意味構造と呼ぶ木構造を生成する．木構造のノードは概念名（記号）と意味関係名からなる．

概念抽出は，意味関係のうち入力言語の文体を担う特殊な関係を概念関係（事実内容を記述する有意味な関係）で置換した概念構造を抽出する．

これらふつの処理は，入力言語内に閉じた処理である．そして概念構造を記述する概念記号（概念名と概念関係）は，入力文の意味内容を記述する記述子であり，各言語共通な意味を代表する．例えば，図1の概念構造中の記号"LOCation"，"TOWER"，"TOKYO"は日・英両言語で同一意味内容をもつ．

目的言語内に閉じた処理としては，文体選択，構文生成，形態素生成のふモジュールがある．この生成処理は，概念記号に対して目的言語内の対応する単語表現を辞書により見い出し，文法的に誤りのない文表現を得る．

文解析・文生成処理は，概念構造をもって基本的に2分される．この概念構造は，各言語の文体の差違を吸収し，事実内容を充分表現できるよう，全ての概念記号が何らかの表方の単語に対応付けられる．即ち，図1例中の"LOCation"に代表される関係概念も"ある"，"存在する"，"の（東京のタワー）"，"in"，"exist"，"place"等の表現と対応付けられる．こうすることにより，トランスファー方式でトランスファー処理の対象となる文体変換・構文変換の多くが不必要となる．

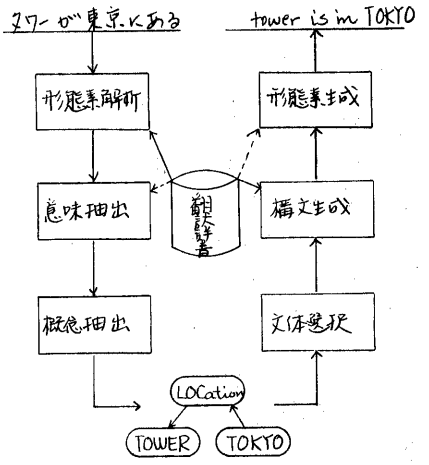

図1．VENUS翻訳手順

2.2 語彙辞書構成の要件

VENUS翻訳システムの処理内容は，翻訳システムやその翻訳方式が異なる場合も何らかの手段で処理すべき事柄である．つまり形態素処理，構文処理，意味・概念処理機能は，多くの翻訳システムの必須機能である．そこで，翻訳用辞書構成に対する要件について VENUS を土台に整理する．

＜異なる言語の対応付け＞

翻訳システム用語彙辞書は，異なる言語間の対応付けを語彙レベルで記述する．VENUSでは，概念記号が仲立ちとなり，日本語の単語と英語単語（より正確には両言語の形態素）を対応付ける．ちなみに，トランスファー方式では，日本語単語と英語単語の対応辞書（トランスファー辞書）がこの対応付けを管理する．この対応は，一般にn対nであり，この対応関係を保持・管理することは，訳語選択の質を向上させるうえでの鍵となる．

とくに，辞書開発後翻訳システムに

組み込んで，実行評価を行う際に誤った対応を検出するとか，そこに新しい語を追加する場合には，対応の明示が必要となる．

辞書が2言語以上の対応を保持する必要がある多言語間翻訳では，この対応管理はより重要な課題となる．つまり，言語Aの単語aから言語Bの単語bへ対応が付けられ，言語Cの単語cから言語Aの単語aへ対応付けられたとき，言語Bの単語bは単語aとcに対して正しい対応を保持していることをどのように検証するか．

<u>〈同義語管理〉</u>

機械翻訳では，訳文・訳語の均質性・・・ひどくまずい訳文や訳語が存在しない・・・が利用者から期待される．複数の同義語が1つの訳語に割り当てられることがある．逆に1つの単語が複数の同義語に対応付けられることもある．とくに単一言語辞書が解析生成の両処理で用いられる場合，上記のn対1，1対nの対応が翻訳辞書で同時に存在し得る．

訳語の統一性・均質性を保持するためには，n対nの同義語対応が明示的に表現でき，その同義語が均質的であるように保障する手段が必要である．とくに，既に登録された単語と同義な単語を辞書登録する場合に以前に付けた訳語との対応を無視すれば，訳語の不統一や不均質が生じる．

<u>〈辞書記述の重複排除〉</u>

膨大な辞書情報を記述するのに，全く同一の内容を重複して記述することは回避すべきである．市販のMT化辞書がそのまま利用できない現在では，質を落とさず記述量を減少させることが望まれる．例えば，多品詞語も見出し語は1回記述すればよく，同一品詞多義語は，見出し，品詞情報は1回の記述で充分である．また，多品詞語でも意味が同じなら，意味記述は1回におさえたい．同品詞・同義語ならば，見出し以外は重複して記述したくない．

辞書記述重複排除は，単に記述労力を軽減する目的より，記述の誤りの最小化を大きな目的とする．大量の語彙に対し，同一内容を何度も記述すると誤りの発生は避けられない．そのためには，初期辞書記述（登録）段階から何らかの方策を用意しなければならない．

<u>〈不連続語彙登録〉</u>

言葉には，不連続な単語（熟語の1部）がある．通常熟語辞書という特別な辞書として開発されたり，不連続語彙の中心語基の付属成分として同語基辞書内に記述される．しかし，処理の観点から視ると，このような単語の同定には，通常の分かち書きした日本語の場合）や派生語認定のための形態素解析（英語の場合）以外の辞書引き動作が必要である．更に，英語不連続単語のように，活用（形態）処理以前と以後両時点で語同定を行う必要が生じることがある．

また，特定語基の付属成分として記述する場合にも，付属成分を入力文中で検索する特別な処理が必要である．

こうした不連続語の辞書記述は，単に記述できるだけでなく，処理手順の簡素化にも何らかの配慮が必要となる．

以上述べた要件の他にも様々な検討課題がある．機械翻訳以外の多様な応用に関し適用可能な内容と，適用可能な構造を持たせる．辞書の管理・維持が容易である．そして，その管理・維持方式がシンプルである．など，辞書記述能力とそれを支える辞書構成方式から，その辞書を機械によって維持・管理する強力な機能やシンプルで使い易い辞書管理システムとのインタフェースまで満足しなければなら

ない要件は多い．

3. 語彙辞書構成

本節では，語彙辞書の構成法とその特徴について述べる．

3.1 三層辞書構成

三層辞書構成は，形態素辞書，構文辞書，意味（概念）辞書の3つの辞書により構成される．各単語の辞書記述も，3種の情報に区分される．三層辞書の物理構成の概略を図2に示す．

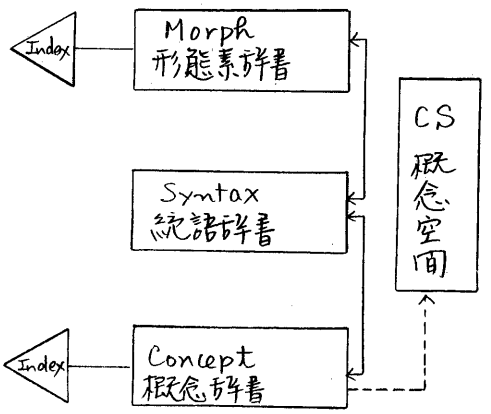

図2. 辞書物理構成

形態素辞書と概念辞書は，各々形態表記と概念記号を見出しとして検索することができる．形態素辞書と構文辞書は互いに双方向のポインターで結合され，構文辞書と概念辞書間も双方向ポインターで連結される．

この三層構成によれば，形態素見出しを与えることにより，形態辞書，構文辞書，概念辞書が検索でき，逆に概念辞書から構文辞書，形態辞書が検索できる．この検索過程は，入力文から概念記号で表現した概念構造を抽出し，概念構造から目的文を生成するTENUSの翻訳過程に論理的に対応する．

なお，1つの形態辞書項目は一般に複数の構文辞書項目に対応し（多品詞語），逆に1つの構文辞書項目も多数の形態素辞書項目に対応する（異表記）．また，1つの構文辞書項目が複数の概念辞書項目に対応したり（多義語），1つの概念辞書項目が複数の構文辞書項目に対応したりする（同義語）．とくに概念辞書項目から形態素辞書を視ると，複数言語にまたがる単語の対応が鮮明に理解できる．図3は，それを模式図で表わす．同図において，M，S，Cはそれぞれ形態辞書項目，構文辞書項目，概念辞書項目であることを示す．例えば，概念記号「REASON」

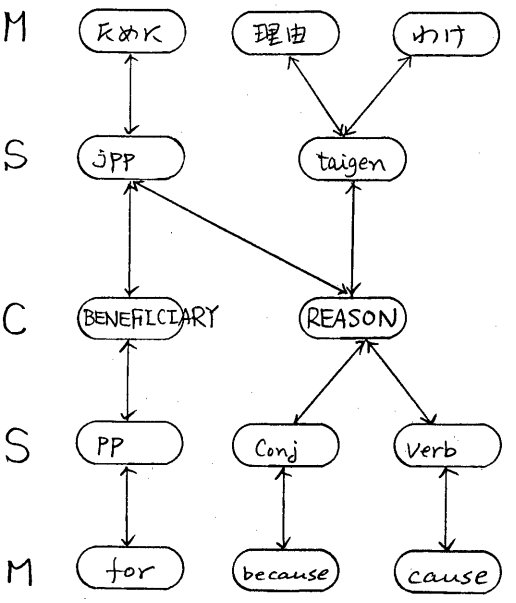

図3. 異言語間の対応模式図

は，構文機能も表記も異なる日本語・英語の表現という対応するかが簡単に読み取れる．

3.2 辞書記述

辞書構成が，3種の辞書を如何に物理的に関連付けるかを規定する．その内容とその構成にどんな意味付けを行う

かは，辞書記述の問題である。

辞書記述に於いてはまず次の制約を設けた。　すなわち形態素辞書項目に対応を持たない構文辞書項目は存在できない。　つまり，構文情報（構文機能）は，表層表記（形態素辞書見出し）に対して記述するということである。この制約は，言語処理の直接の対象である形態素に対し辞書記述を行うという自然なものである。

形態辞書見出しは，その表記と発音（読み）からなる。　異表記・同一単語（形態素）に対しては唯1の構文辞書項目を与える。

1形態素に関し，複数の構文機能（例えば品詞）が存在する場合は，複数構文辞書項目に対し唯1の形態辞書項目を与える。

構文辞書と形態辞書記述の関係は，構文辞書と概念辞書記述の関係と相似である。　つまり，多義語，同義語（言語を越えた同義語も含む）に対する概念辞書項目，構文辞書項目は各々唯一の構文辞書項目，概念辞書項目に対応させる。

このような3種の辞書項目により1つの語彙を記述する手法に於いては，

a) 記述重複の回避と，それによる記述誤りの局所化が図れる。
b) 概念辞書を中核とした同義語の明示的管理と，それによる同義語どうしの記述内容の矛盾排除が図れる。
c) 概念辞書を中核とした対訳同義語管理と，それによる訳語品質劣化の防止が図れる。
d) 概念辞書を中核とした多数言語にまたがった語彙対応管理の一元化が図れる。

などの特徴がある。

さらに，

e) 各辞書の機能を細分化することにより，簡単に熟語辞書を記述でき，通常の語彙辞書記述と同じ枠組で管理できる。

通常熟語というといくつかの種類がある。　まず，複合語と呼ぶ連続したものと複数の語からなる不連続なものに2分される。　そして，各々は活用するものと不活用のものに2分される。三層辞書構成に於いて，不連続・活用型熟語も図4に示すように記述できる。

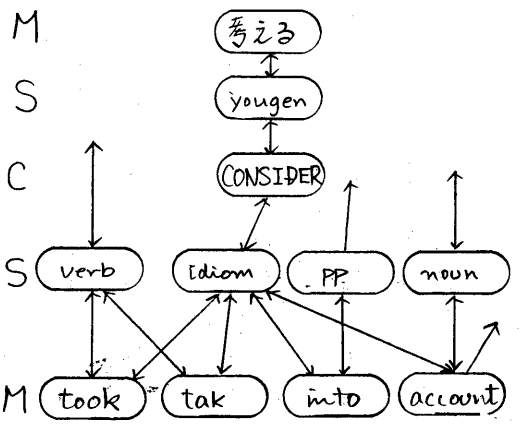

図4.　不連続単語の記述

同図に於いて，構文辞書Sの項目「idiom」は，構文辞書機能を拡張した熟語保持項目である。　この項目には，「考慮する」という意味の熟語が記述されている。　つまり，以下の特性が言える。"take"，"took"，"into"，"account"などが熟語要素であると同時に，各々独立した構

文辞書を持つ．　故に，この形態辞書を用いて分かち書きや形態素解析時に熟語候補を選びぬくことができる．また熟語認定には，「idiom」項目中に条件記述することにより対処できる．即ち，熟語認定のために再度熟語辞書をアクセスしたり，熟語付属成分を入力文中で検索したりといった無駄が省ける．

4.　辞書管理

辞書管理の重要な機能は，辞書検索機能と，辞書項目の登録および削除などの更新機能である．

辞書検索は，形態素見出しと概念記号をB-treeインデクスとして実現され，他の辞書項目へは，関連辞書項目へのポインタをたどることにより到達できる．

辞書更新には，辞書管理エディターと，辞書登録ソフトウエアツールを用意している．　辞書登録ソフトウエアツールは，書式に従って書かれた各辞書項目を一括登録する機能を持つ．それに対し，辞書管理エディターはオンライン端末により会話的に登録・削除を行うために用意した．　図5に，同エディターの概念図を示す．

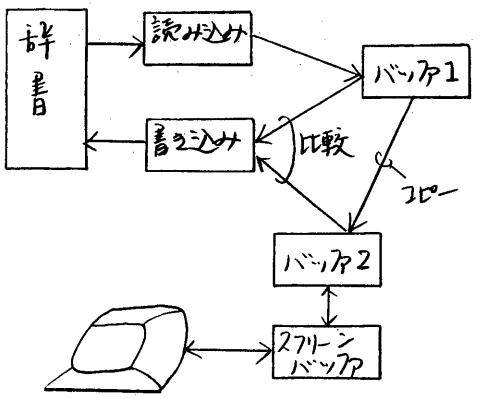

図5．　辞書管理エディター概念図

辞書管理エディターは，基本的に形態素見出しと概念見出し（概念記号）の指定による辞書アクセスを提供する．以後は，どの辞書項目にアクセスするかを，ポインタで連結された項目のうち画面に表示されたメニューに従い利用者が指定する．

指定された見出しが存在しないときは，登録モードが，見出しが既に存在するときは更新モードが自動的に選択される．　エディター画面のメニューに従いアクセスした全ての辞書項目はバッファ1に貯えられ，エディット結果を保持するバッファ2と比較した後更新される．　また，エディター画面はスクリーン機能を持ち，スクリーンバッファにより管理される．

辞書管理エディターは，それを用いることで，ポインターで連結されたどの辞書項目へもアクセスでき，アクセスした辞書項目に関連する新規辞書項目の登録が可能になっている．　こうした強力な機能は，一旦開発した辞書内容を比較し記述の誤りを発見訂正するための有力な支援となる．

辞書管理エディターの直感的理解のため，次頁図6．a～dにエディター端末（PC9800）画面を例示する．　同図は，概念記号「PROCESS」を指定して，更新モードで各種辞書項目をアクセスするときの画面イメージである．

図6.aは，概念記号「PROCESS」を指定したときの画面であり，「PROCESS」に対応する形態素表記が同時に表示される．

図6.bは，前画面より形態辞書へのアクセスを指示した結果であり，形態素の選択を待つ．

図6.cは，前画面で形態素「処理」を選択し，その内容が表示された．

PROCESS	CCS. (AGT, OBJ, INS)
処理 プロセス 加工 加熱処理 process	

in Conceptual Screen

block number 16289
Index number -1

RUP RDOWN SASK SKIP SAVPTon CPY BF LPT Lf

図6.a 概念許書項目画面

表層								
接続	A				F			
	B				G			
	C				H			
	D				I			
	E				J			

in J. Morph. Screen

KEY input or NUMBER input (space/figure) =
0 = 処理 1 = プロセス 2 = 加工 3 = 加熱処理

MORPH SYNTA CONCE SKIP SAVE LTRL HALF

図6.b 表層許書項目選択画面

表層	処理							
接続	A	204 234		1		F		
	B					G		
	C					H		
	D					I		
	E					J		

in J. Morph. Screen
block number 16291

CP = PROCESS

図6.C 表層許書項目画面

Syntax 1	Conceptual Primitive
	PROCESS
	Morph
	0. 処理

in Syntax Screen
block number 16288

CP = PROCESS -> Mono = 処理 ->

図6.d 構文許書項目画面

図6.dは，前画面の「処理」に対する構文許書項目の表示である．

このような画面を介し利用者は三層許書へのインタラクティブなアクセスを行うことができる． この場合，もし形態素が複数の構文許書項目を持つとき，このアクセスメソッドはいわゆる類語検索機能に似た役割も果し得る．

また，この管理エディターは新規登録作業を円滑にするための機能としてAKO（A-Kind-Of）機能をもつ．AKO機能では語彙登録時に全ての情報を全く新しく記述する代わりに，多くの情報を共有する既登録許書項目を指定しその内容を編集する。こうして新しい語彙に対する許書登録を容易にする．

5. おわりに

日英機械翻訳システムの翻訳許書構成とその管理システムについて述べた。構成の特徴は以下の2点である．

a. 形態許書・構文許書・概念許書からなる三層構造を導入した．

b. 概念許書を異なった言語間を連結するものとして中心に置いた．

この2つの特徴から以下の機能が実現できた．

1. 類語検索
2. 同義語検索
3. 語彙許書・熟語許書の統合
4. 重複記述の排除とそれによる冗長の最小化

また，この許書構成により，概念許書（概念記号）を中核として多言語許

書を統一的に管理することができる。同様に，同一意味内容をもつ語群を一括管理することにより，誤語品質の劣化を防止する手だてを提供する。

この構成をもつ辞書データベースはVENUS翻訳システムに用いられているが，いくつかの問題点も明らかになった。それは，辞書が3種類に分割されたことにより，利用者がそれを明確に意識する必要が生まれたこと，そのため，利用者の負担が増加しているように感じられることである。実際，画面や辞書フォームに従って辞書登録する際，各々の辞書の対応付けに無駄を利用者が感じることがある。これはあくまで心理的な事柄であるが，この辞書データベースの利用をより容易にするための方策を考える必要がある。

また，概念記号は利用者が管理し登録するものである。それが形態素表記群を代表するのに適切であることを保障する必要がある。これは，とくに知的作業に属し，実行にはコストがかかる。もし，概念記号を単に両言語の語句の一対一対応を保持するための道具と仮して，1つの対訳のインデクスとするならば，概念記号の付加作業は全く機械的に行える。しかしながらそうすると，辞書の多言語対応能力が減少し初期要件を満足することができない。同義語(多言語にまたがる)を管理する概念記号付与作業を支援する方策を検討する必要がある。

今後このような問題を解決するため，辞書管理システムの運用を通して実践的方策を考えていきたい。

最後に，本研究開発の機会とそれに対する支援を与えられたCoCシステム研究所・メディアテクノロジ研究部千葉部長ならびに山本課長に深謝します。同部荻野主任には，管理システム設計に協力していただきました，ここに感謝の意を表します。

[参考文献]

1.) 中村順一，他，「Muプロジェクトにおける辞書の運用方式－日英交換辞書と英語生成辞書－」，「自然言語処理技術」シンポジウム予稿，1984，11

2.) 荻野，浜田，「自然言語処理と辞書管理」，情処大会37回，1984.9

3.) OKAJIMA，他，「Lexicon Structure for Machine Translation - An Example from English-into-Japanese Translator ATHENE -」，1983 International Conference on Text Processing with a Large Character Set