解　説

辞書の構成と機械翻訳†

辻　井　潤　一††

1. はじめに

機械翻訳システム（MT システム）は，人間の翻訳家が持っている言語使用に関する知識を集積した知識情報処理システムの典型である．MT における知識は，言語についての一般的な知識を表現した文法規則と，単語に個別な語彙的知識を集積した辞書とに分けられる．この2つにどのような知識を投入するかでシステムの能力が決まる．しかしながら，実際には，辞書に品詞情報しかなければ，文法規則を精密化するのにも限界があるように，辞書中の知識がシステムの枠を規定することになり，意味処理，知識処理の導入を含む処理の高度化は，必然的に豊富な辞書記述を要求することになる．一方，辞書記述は少なくとも数万から数十万の単語を対象とすること，また，対象分野を変えるごとに単語の新規登録が必要なこと，などの理由から辞書記述ができるだけ簡便であることも要求される．このようなことから，辞書をどうするかは，MT 技術における最大の課題となっている[1]．

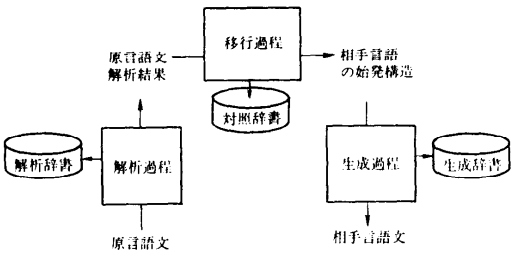
本稿では，われわれが科技庁の Mu-プロジェクト[2]（以下では，Mu）の辞書を設計した経験をもとに，MT における辞書の種類，辞書に記述すべき情報，辞書の作成・管理の問題，などについて展望する．

2. 辞書の種類

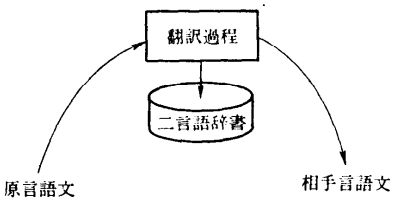
2.1 翻訳方式と辞書の種類

MT の方式としては，現在，トランスファ方式，直接方式，中間言語方式が提案されているが，これらの方式ごとに使用される辞書の種類に違いが見られる．

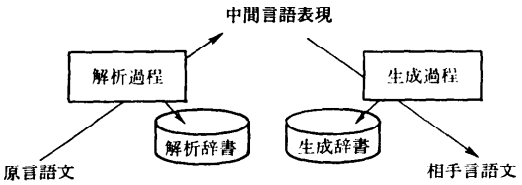
トランスファ方式は，翻訳を解析，移行，生成の3段階に分けて実行する．各段階は，それぞれ原言語解析辞書，両言語対照辞書，相手言語生成辞書を使用する（図-1(a)）．解析辞書，生成辞書はそれぞれの言語の範囲内で考えれば良いので単一言語辞書，これに対して，対照辞書は二言語辞書と呼ばれる．



(a)　トランスファ方式と辞書



(b)　直接方式と辞書



(c)　中間言語方式と辞書

図-1　翻訳方式と辞書

直接方式は，トランスファ方式とは異なり，原言語を解析した結果を直接相手言語の語彙と構造で表現する．したがって，この方式では翻訳の全過程を通じて，二言語辞書だけが用意される（図-1(b)）．直接方式においても数種類の辞書を用いることがあるが，その区別は，単一言語辞書，二言語辞書とは別である．

これに対して，中間言語方式では，翻訳過程を原言語の解析と相手言語の生成の2段階に分けて実行し，両過程は言語に依存しない中間言語での表現を介して結ばれる．したがって，解析辞書と生成辞書の2つの辞書を使用するが，これらは，それぞれ原言語の語彙か

† The Roles of Dictionaries in Machine Translation by Junichi TSUJII (Department of Electrical Eng., Kyoto University).

†† 京都大学工学部電気工学第2教室

ら中間言語の語彙(概念語)への写像，および，概念語から相手言語の語彙への写像を規定する（図-1(c)）．すべての言語のすべての語彙の意味を完全に記述できる中性的な中間言語が設定できれば，この方式では，解析辞書，生成辞書ともに単一言語辞書となる．

2.2 運用面からの辞書の区分

実際のMTシステムでは，翻訳処理での使われ方の相違や，処理効率，システム運用の形態，辞書データの管理などへの配慮から，単一言語辞書と二言語辞書の区分の他に，さまざまな種類の辞書区別を設けている．現在，いくつかのシステムで設けられている辞書の区分を以下に列挙する．

［処理効率のための区分］高頻度の辞書とそれ以外の単語辞書を区別し，前者を主記憶上に常駐させることによって辞書引きの効率を上げる．

［処理階層からの区分］原言語の解析過程は，さらに形態素処理，統語処理，意味処理などに区分され，それぞれの段階で必要とされる情報が異なる．各段階での辞書を別々に管理することによって，たとえば，形態素辞書はすべての分野に共通に使えるといった記述のモジュラリティが確保できる．特に，知識処理用の情報は，対象分野への依存度が高いので，一般の言語情報用辞書とは独立に管理する必要がある[3]（図-2）．

［複合表現辞書］MTにおいては，'on the other hand' のように，部分の意味（訳）から全体の意味（訳）が復元できないもの，'developing country'（発展途上国）のように複合表現として登録しておけば，他の解釈や訳語の可能性がなくなるもの（developing country には，『国を発展させること』の解釈もある），などを1つのまとまりとして取り扱うことが多い．このような複合表現は，形態素解析など多くの処理で特別扱いする必要があることから別に管理されることがある[4]．

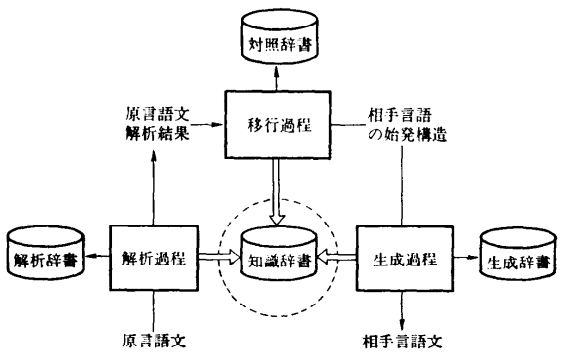


図-2 知識辞書と言語辞書

［専門分野別辞書，ユーザ用辞書］対象分野によって使用される単語が変化したり，表面上同じ単語の意味や訳語が変化することから，分野ごとに辞書を管理することがある．ただ，特定分野のテキスト中にも他分野の用語が混在するのが普通であり，専門用語の範囲を限定することは困難である．逆に，同じ分野の同一用語でも各ユーザごとに異なった訳語を使う場合もあり，この場合には，専門分野別辞書よりもユーザ別辞書となる．

3. 辞書の記述内容

実際のシステムでは，システムの形態に応じて独自の形式を持った数種類の辞書が使われるが，表面的な記述上の相違を捨象すると，MT用辞書に記述すべき情報をある程度整理することができる．本節では，語数が多く辞書の作成の際に最も大きな問題となる動詞と名詞を中心に，どのような知識が単語ごとに辞書に記述されるべきかについて述べる．

3.1 単一言語辞書

表-1 に，Mu の日本語動詞辞書での記述項目を示す[5]．注）に示したように，1つの動詞にはその意味に応じていくつかの用法があり，動詞の意味分類などの情報は，用法ごとに記述される．以下に，代表的な記述項目について説明を補足する．

［関連語］類義語，同義語，反対語などを記述する．動詞の場合には，他動詞から自動詞への，あるいは，その逆のリンクもこの欄に記述される．この記述は，生成過程において，自他の切り換え（英日の場合には，しばしば必要）を行ったりするのに使われる．より柔軟な文生成を行うためには，S1（研究する）->『研究者』，S2（攻撃する）->『攻撃対象』といったメルチューク・岡本らが語彙関数で整理した単語間の関係も，この欄に記述すべきである[6]．派生用接辞（者，機，器，など）やそれと同等の意味機能を持つ名詞（装置，システム，など）の使い分けが単語固有であることから，語彙ごとにこの種の関係が指定されていることが望ましい．

［動詞意味分類］動詞意味分類としては，例えば，高松らによって提案されたものがある[7]（図-3）．高松は，これを使って，動詞と格助詞の共起を予測したり，その深層での解釈を行う規則を定式化している．しかし，多くの動詞を

表-1 日本語動詞についての情報

見出し語		見出し語の標準つづり
	語尾字数	
	漢字部	
	語基	複合語の場合，その構成単語
	読み	
異形語		異つづり
派生語		その語からの派生語と派生の意味的関係
関連語		類義語，反対語，自他の対応，可能動詞
形態素情報		活用型
分野コード		翻訳対象分野（現在はすべて電気分野）
構文品詞		
用法ごとの情報	ID	複数個の格枠がある場合の識別記号
	変換見出し	二言語辞書での見出し語
	アスペクト	動詞のアスペクト素性による分類
	態変換	受身，使役の場合の表層CFの変化
	意志	意志性による動詞の分類
	意味分類	記述動詞，移動動詞などの分類
	シソーラス	シソーラスでの登録語（現在使用せず）
	格枠	表-2 参照
	共起情報	副詞など特に共起しやすい単語

注）用法ごとの情報は1つの動詞について，複数個つき得る．

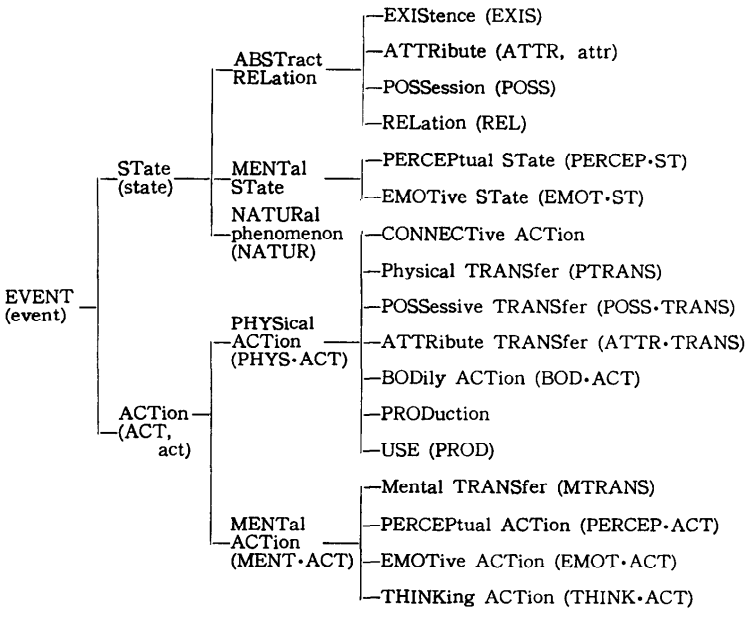


図-3 動詞の意味分類

分類するには基準が明確でないこと，同じ分類の動詞でも異なる格助詞を要求するものがあることから，格助詞との共起などをすべて動詞の意味分類から規則化することには無理がある．このような関係は，個々の動詞ごとに辞書に与えておく方が現時点では無難である．ただし，動詞ごとの記述が冗長となる自由格の解釈，あるいは，文接続表現の意味解釈などには，ある種の動詞意味分類が有効であろう．

Muでは，図-3のような詳細な意味分類ではなく，動詞のアスペクト素性による分類(瞬間，継続，など)と意志素性による分類（意志，無意志，準意志）とを用いている．アスペクト分類は，日本語の『ている』，『てくる』などの解釈を行い，それを英語の進行形や完了形に変換する場合に必要となる．また，この程度の意味素性であっても解析過程で自由格要素や文接続の意味解釈を行うのにかなり役に立つ[8),9)]．

［格パターン］2つの言語における構造の対応は，

read X<-->X を読む

mary X<-->X と結婚する

のように，直接的ではない．このような対応をとるために，Muをはじめ現在のMTシステムでは，原言語文の意味を深層格構造で表現し，これを仲介にして相手言語での統語構造を決定することが多い(図-4)．この場合，解析過程では，表層の格表現を深層の格解釈に，生成過程では，深層格関係を表層の格表現に，それぞれ写像しなければならない．この写像が単語個別であること，また，動詞が表層上どのような格枠（Case Frame-- 以下では，CF）を持つかが単語個別であることから，動詞ごとにとりえるCFとその深層格解釈との写像関係を記述しておく必要がある．

動詞のCFには，一般的に，格助詞や前置詞との共起制限のほかに，各格要素ごとに，

（i） どのような統語的構造をとれるか（英語の場合，名詞句，That-補文，Wh-節，Ing形の補文，など：日本語の場合，こと-補文，の-補文，連用形，名詞句など）

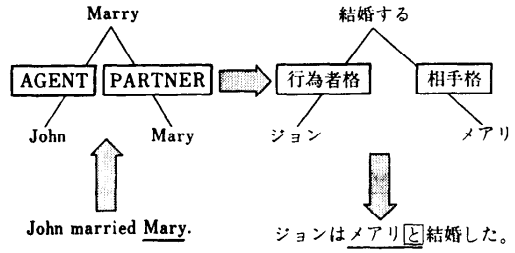


図-4 深層格構造

(ii) 格要素が満足すべき意味的制限

が記述される．この種の情報は，解析の曖昧さを減少させるだけでなく，適切な補文形式を選択するなど文生成過程でも必要となる．

[必須格と任意格] 同じ格助詞『と』でも，行為に必須的に付随する要素をマークする場合(『Xと結婚する』)と，任意的な要素をマークする場合(『Xと行く』)とがある．解析過程において，名詞句の係先に複数の動詞が可能な場合，必須な関係を優先することで正しい解釈が得られることが多い，などの理由から，これらの格関係を区別しておくことが望ましい．

Mu の動詞の格枠の記述形式を表-2 に示す．日本語動詞の CF については石綿の分類[10]，英語については，Longman や Hornby などの人間用辞書の動詞分類コードや，CF と深層格解釈の関係を大量の単語に対して調べたインディアナ大学の単語リスト[11),12)]，などが参考になる．

[態変換パターン] 動詞の標準的な CF は，さまざまな要因によって変形を受ける．その一部は文法規則だけで処理できるが，語彙的知識を必要とするものも多い．例えば，受動態による変化は，『彼からその事を知らされた』，『彼に殴られた』のように動詞によって異なるために，自然な文の生成には，語彙的知識が必要となる．Mu では，受動態，使役態など語に依存する変形をパターン化し，各動詞ごとに指定している．

名詞についても，適切な数詞(つ，人，個，など)や序数詞(第〜回，代，など)は名詞ごとに変化すること，動詞の CF と同様に，深層格を表現するのに特定の前置詞や格助詞を要求したり，修飾句に特定の構文形(例えば，N for one's〜ing)を採る名詞があること，『昨日』・『少数』・『安全』のように構文的・意味的には副詞・連体詞・形容動詞とよく似た動きをする名詞があることなど，必要な語彙的知識は数多くある．また，動詞のCF中に格要素の意味的制限を記述するためには，名詞の意味的性質を記述する手段が必要となるが，Mu では，58個の意味マーカを使って名詞の意味的性質を表現している[2)]．

表-2 動詞の格枠の記述

表層格表現	日本語：とりえる格助詞のリスト 英語：主語・第一目的語などの表層での位置，および，前置詞句の場合には特定の前置詞
深層格解釈	深層格
統語カテゴリ	日本語：こと-補文，の-補文，連用形，など 英語：ING 形の補文，To-不定詞句，That-補文，Wh-補文，など
補文統語形	補文のアスペクトやモーダルの指定
意味的制限	名詞意味マーカのリスト，または，(その名詞と動詞の共起が慣用的である場合には)特定の名詞
任意／必須	その格が必須的か任意的か

注) この他，動詞のCFには特によく共起する副詞や形容動詞連用形が指定される．

3.2 二言語辞書

3.2.1 記述の形式

二言語間の語彙対応が一般的な規則で処理できない以上，最低限，この対応を単語ごとに持つ必要があるが，二言語間の語彙対応は一般に多対多であり，単純ではない．一般には，原言語文における語の使用環境を調べて，最も適切な相手言語での訳出形を選択しなければならないために，この対応の記述は，条件部とそれが満足された場合の相手言語での表現という条件付きの対応となる．また，

(1) 訳出形の選択のためには，できるだけ広い範囲の使用環境を柔軟に検査できる必要があり，条件部での豊富な表現能力が必要となる

(2) 原言語の単語が相手言語でも単語に対応する保証はなく，表現対表現の対応が記述できる必要がある

などの理由から，この対応の記述は，いわゆる辞書的なデータではなく，手続きになることが多い．

実際，トランスファ方式をとったカナダの TAUM-AVIATION でも，二言語辞書記述用のプログラム言語 LEXTRA を使って手続き的に記述しているほか[13)]，田中の Active-Dictionary[14)]，シストランの CLS (Conditional Limited Semantics)[4)] など手続き的な記述が使われている．Mu においても，二言語辞書中の任意の語に対して直接 GRADE[15)] の木構造変換規則を定義することを許している．

しかしながら，TAUM-AVIATION の最大の難点

が辞書の作成コストにあったように，記述の柔軟性は逆に辞書作成の困難さに繋がる．現実に大量の語彙の記述を行うためには，なるべく多くの語彙に対する記述が一応可能な範囲で，ある程度制限された辞書の記述形式を設定することが必要になる．大量語彙用の制限された記述形式を設定しておくことは，辞書の共用性を高める上でも重要である (4.2節)．このことは二言語辞書だけでなく，単一言語辞書にも言えることであるが，対応が複雑で手続き的記述になりやすい二言語辞書の場合には，特に注意する必要がある．

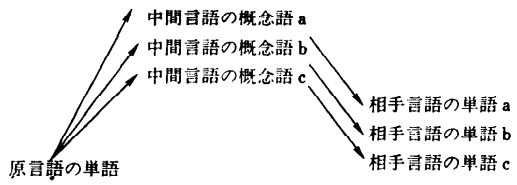
以下の各節では，日本語と英語の多くの単語についての対応をとるために，このような制限された記述形式が最低限どのような程度の記述能力を持つべきかについて考える．

3.2.2　語の共起と訳語選択の条件

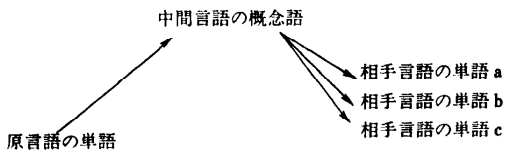
原言語内での明らかな多義性は，単一言語辞書ですでに異なった用法として分離されており，二言語辞書で訳し分けの条件を記述する必要はない．しかし，単一言語内の多義なのか，相手言語の語彙との関係で訳し分けが生じるのかが微妙な場合も多い[16] (表-3)．

二言語辞書を認めない中間言語方式では，これらを原言語内の多義とし中間言語で区別するか (図-5(a))，あるいは，意味上の差はないとして，中間言語から相手言語を生成する段階で具体的な語の選択を行うことになる (図-5(b)) が，いずれの場合でも，以下で議論する二言語辞書に必要な情報は，解析辞書か生成辞書のいずれかに記述される必要がある．

表-3 の(a)～(d)例は，いずれも述語の訳語がそれに結び付いた格要素によって選択される例である．『語の使用環境』の最も局所的なものが，このような2



(a)　原言語での多義



(b)　相手言語での表現上の慣習

図-5　中間言語での訳し分け

表-3　共起と訳し分け

a	'heavy rain' 'heavy bags'	激しい 重い
b	'play tennis' 'play piano' 'play drum	する ひく たたく
c	'cut a hedge' 'cut a tree'	刈る 切る
d	『速度を上げる』 『効率を上げる』 『水面を上げる』	increase, improve improve raise
e	『最近の進歩』 『最近の製品』	recent new
f	『電流の強さ』 『材料の強さ』	intensity strength

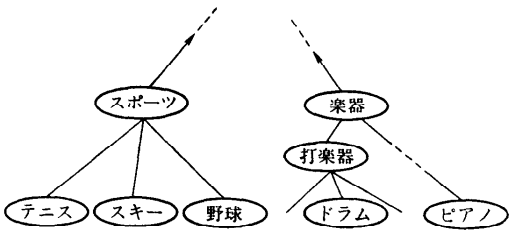


図-6　概念階層

つの単語の共起である．辞書記述を行う場合には，格要素のどのような『性質』が述語の訳語選択に関与したと考えるかが問題となる．次の3つの立場がある．

(1)　『激しい雨』<->'heavy rain' のような対応は，両言語での慣習の問題であって規則化することはできないとし，2語の結び付きに対して直接訳語を与える立場*.『風邪をひく』-->'catch cold' の対応なども，この立場からとらえることが多い．

(2)　訳し分けには，現実世界の知識が関与していると考え，概念間の階層関係を導入する立場．(b)の訳し分けには，図-6 の概念階層が関与していると考える．

(3)　(2)と同様に現実世界に関する知識を援用するが，これを素性で捉える立場．(c)の訳し分けには，日本語では『密生した固いもの』を 'cut' する場合には，『切る』よりも『刈る』の方が自然である，といった規則性があるが，このように，語が指示する現実世界の『もの』の性質の差が訳し分けに関与する場

* heavy rain のように形容詞の名詞修飾も深層格構造の段階では，関係節構造として通常の述語と格要素の関係として取扱うことが多い．

合には,『もの』それ自体を分類する概念階層より,性質の共通性を捉える素性の方が適している.

(2)と(3)は,いずれも,膨大な知識の整理が必要となり,現時点ではあまり期待できない.したがって,現時点では,かなり粗い概念階層や素性で訳し分けを行い,それで訳し分けできないものは(1)の語同士の共起を直接辞書に記述する方式がとられる.Muでは,(1)と(3)を使い,(3)による訳し分け条件を,前述の名詞の意味マーカを使って動詞の辞書に,また,(1)の訳し分け条件を,述語の訳し分けを要求する名詞の変換辞書に,それぞれ記述している.

このような共起関係による訳し分けは,例(f)のように名詞にも必要である.

3.2.3 訳語選択の要因

相手言語での表現の選択には,前述の語の共起の他にさまざまな要因が関係している.以下にいくつかの要因とその例を示す.

[係り先の動詞句の素性]『簡単に』は,副詞 'briefly', 'easily', 'simply' などに訳されるが,この訳し分けは述語の意味的な性質(述べる,説明する,などの一連の動詞に共通する素性--Muではこれらを記述動詞と分類している)や意志性などを調べることによって,多くの場合,訳し分けられる.

[原言語文での単語の意味]サ変名詞は,『〜すること』の他に『〜した結果のもの』,『〜する人』など,種々の意味を表現しえる(たとえば,監督).通常の名詞でも,『豚』には『動物としての豚』とそれから派生した『食物としての豚肉』の両義があり,それぞれの意味に応じた訳語を選択する必要がある.

[語の意味的役割]『木の成長』と『木の机』の『木』は,それぞれ 'tree' と 'wooden' とに訳し分けなければならない.この訳し分け条件には,『木』が『成長』や『机』に対して持つ深層格関係が関係している.『安全に->safety (安全に注意する), safely (安全に運転する)』の訳し分けも,述語に対する深層格が関係している.

[述語の素性]『死んでいる -> dead : 死ぬ-> die』の対応では,述語のアスペクト素性(状態,未完)によって単語を選択する必要がある.

[統語的な解釈]『'be educated'-> 教育される,教育がある』の対応は,解析の段階で受動態と解釈されたか,過去分詞の形容詞的用法と解釈されたか,によって訳し分けられる.過去分詞に形容詞的用法を一般的に認めるとすると,動詞の二言語辞書に形容詞的に使われた場合の記述が必要である.

[述語の態] 'I was given a book'->『私は本を貰った』のように,英語での態が日本語では動詞そのものの中に入ってしまう場合がある.

3.2.4 語対表現,表現対語の対応

日英のように語族の異なる言語間の対応においては,単語の対応が必ずしも語対語とはならない.この場合には,条件記述だけでなく,変換後の記述も複雑になる.ここでも,制限された記述形式においても最低限取り扱えるようにしておくべき現象について考える.次のような現象は極めて頻繁に見られる.

[格要素と述語の組合せ]『効率が高い<->efficient』は,日本語の複合表現が英語で1語になる例,『試作する<->develope 〜 on trial bases』は,その逆である.『ダイアルをまわす<->dial』のように,日本語の目的格と述語が全体として英語の動詞になる例は多い.

[派生語尾] 日本語と英語では,言語として持っている派生の機構が異なっていること,また,派生語尾を付けた新たな単語が実際に存在するかどうかは単語ごとにことなるため,一般的な規則化はむつかしく,語彙的な知識が必要になる(『離散化する<->digitize』『精密化する<->make A precise』).一方の言語だけに派生語がある場合には,二言語辞書で単語対表現の対応をとる必要がある.

[構造の変換]『注意して<->carefully』のように,日本語のある種の動詞が格要素をとらずに従属節を作った場合,英語では副詞や副詞句になることがある.どのような動詞が副詞句になるか,また,どのような形式の副詞句になるかは,言語対に依存する.

[英語形容詞と日本語の複合表現] 日本語と英語では,日本語の形容詞の数が圧倒的に少ない.したがって,英語の形容詞が日本語では複合表現になることが多い.前述の『効率が高い<->efficient』の他に,『fragile <-> こわれ易い』,『solvable <-> 解決できる』,『paralle<->平行した』,などさまざまな対応がある.

4. 辞書の作成と管理

データ・ベースは,一般に,原データの作成,投入の段階を経て,管理・使用の段階に入る.辞書も一種の大規模データ・ベースであるが,通常のデータ・ベースがデータの作成に大きな困難をともなわないのに対して,辞書の場合には作成時に最大の難関がある.また,この作成時の難関が,データ投入と管理にも大き

な影響を与える．本章では，辞書の作成の問題を中心に，辞書データ・ベースの管理，運用の問題を考える．

4.1 辞書記述の基準

辞書作成時の困難さは，記述基準の設定の困難さにある．自然言語理解システムを研究レベルで行う場合には，数百語の辞書を1人の研究者が作成し，その辞書を使う文法も自ら作るのが普通である．この場合には，辞書記述の基準は問題にならない．研究の進展に応じて，また，対象テキストに応じて，記述を追加していけば良い．しかし，数万から数十万語の辞書は，この方法では作れない．作業者間での記述のユレが生じないように，明確な基準を設定しておかなければならない．また，明確な基準は，作業者の迷いという作業能率を左右する要因や，作成された辞書の管理や記述の網羅性，といった重要な問題とも深く関係している．以下に，Mu で問題となった記述基準の主なものを挙げる．

[単語の表記] 日本語では，送り仮名やカナ表記での長音表記，異文字種による表記など，複数の表記を持つ単語が多い．単語同士の共起を直接辞書中に書き込む場合，どの表記を標準とするかを決めておく必要がある．英語の大文字，小文字の区別やハイフンなども同様である (on-line, on line, online).

[単語の単位] 複数の語からなる表現を辞書に登録する場合，どの範囲までを登録するかが問題となる．接辞や派生語尾(重い->重さ，重み：好む->好ましい)まで含んだ単語を登録するか，などの基準も必要である．

[品詞の設定]『安全』を名詞とするか，形容動詞とするか，あるいは2義性があるとするか，名詞と形容動詞を区別しない品詞体系を採用するか，などの問題がある．また，『従来』,『将来』のように名詞と副詞の判断に迷う場合もある．

[意味用法の認定]『プログラムを走らせる』の『走る』と『汽車が走る』の『走る』とは別の用法であるとするのか，あるいは，同じ用法とするか．また，『合図を送る』と『手紙を送る』の『送る』はどうか，など多分に作業者の主観による．用法を細分する立場に立つと，1つの動詞に際限なく多くの多義性を認めることになり，解析時点の処理が難しくなる．また，用法を細分しないと格要素との意味的な共起制限が甘くなって，どのような名詞とも共起できることになる．また，名詞にも表-4のようにさまざまな観点からの『意味』がある．これらを多義と見るかどうかも，意味マーカを付与したり，概念階層を作る場合に問題となる．

表-4 名詞と多義性

学校	学校を建てる	建造物
	学校で会う	場所
	学校が禁止している	機関
	学校を出ている (学歴がある)	慣用的
	学校に行っている (学生である)	慣用的

[格関係の認定] これまでに開発されてきたシステムごとに格の設定に相違があるように，格の概念は安定したものではない[17),18)]．個々の格の認定基準を明確にしておかないと，作業者の迷いや辞書記述に一貫性がなくなる．また，日本語の場合には，必須格と任意格を区別するための基準も必要である．

[意味分類の付与基準] 意味マーカの付与は，最も基準設定が難しく，作業コストの大きい部分である．豊富な実例とテストの手段がないと作業できなくなる．

[対象分野への依存の程度] あまりに一般的・網羅的な辞書は逆に文法処理への負担が大きくなる．対象分野での使用頻度の低い用法は入っていない方が良い．一方，科学技術分野では，日常の用法にはない語の使い方も現れる．どの範囲までの用法を辞書に入れるかは，結局個々の単語ごとに作業者が判断せざるを得ない．このことによる作業能率の低下，辞書記述の網羅性の低下などは，システムを大規模化したり，他分野へ移行したりする際の障害になる．

4.2 処理からの独立性

人工知能の分野では，フレームなどのように宣言的な知識記述の中に手続き的記述を埋め込むことが一般的であり，自然言語処理においても，辞書中に手続き的な記述を付加して処理の柔軟性を得ることも多い．しかし，この手法の乱用は，作成された辞書がシステム依存的になり文法やシステムを熟知している人間でないと作成できなくなるという欠点につながる．また，少数の語彙が対象の場合には，個別処理が必要な単語が1語だけであっても，大量語彙を対象にした場合には，かなり多くの語彙が同じような手続き付加を必要とすることも多い．このような場合には，それら一連の単語に共通な言語的性質を発見し，どのような基準とテストを使ってその種の単語を見付けるかを整理しなければならない．手続き付加の乱用は，このような語彙的知識の整理段階をバイパスすることにな

る．前節のような辞書記述の基準設定のためには，処理の観点からではなく，言語的な観点からの知識の整理が不可欠である．

現時点では，辞書記述のすべてについて客観的な言語学的意味づけと記述基準を与えたり，単語個別の特殊処理を完全に排除したりすることは不可能である．システムを現実に動作させるためには，言語学的な意味が不明瞭な（システム依存な）記述が混在したり，手続き的な部分があることを避けることはできない．これらをどのように調整して，システム的に組み上げるかが，現実の問題としては重要である．

Mu では，制限された記述形式を使って大量語彙の記述を宣言的に行った，管理の対象となる辞書データ・ベースと，これを変換して，実際の翻訳処理に使う処理用の辞書の２段階をもうけること，また，記述形式に記入される項目についてはできるだけ客観的な基準を設けることで，この問題に対処している．また，辞書データ・ベースの管理対象とはならないが，処理用辞書を直接 GRADE 規則で与えることもできるようにして，処理の柔軟性を保っている[19]（図-7）．

4.3 辞書の作成と管理からみた翻訳方式

直接方式，トランスファ方式，中間言語方式では，それぞれの方式に応じて作成・管理しなければならない辞書の種類が変わる．辞書の作成と管理の観点から見ると，各方式ごとにそれぞれ利点と欠点を持っている．以下では，辞書の作成と管理の側面からこの３つの方式の特徴を考える．

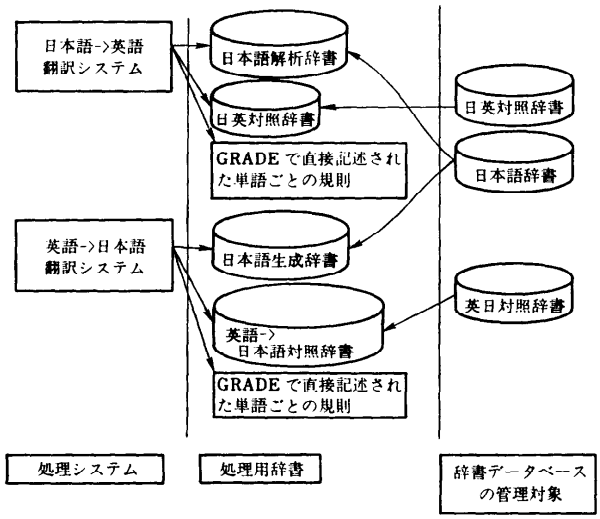


図-7 Mu における辞書システム

[直接方式] この方式で作成・管理しなければならない辞書は二言語辞書の一種類であり，異なる辞書間の整合性を考える必要がないこと，単語の意味の分離なども言語対に依存してできる点では，作成・管理ともに容易となる．しかし，(1)辞書記述に際しては，原言語と相手言語を同時に考えなければならないこと，(2)単一言語辞書がないため，相手言語の単語に関する語彙的知識を重複して記述しなければならない点で，作成時の負担が増大する．また，(3)原言語と相手言語の知識のモデュラリティが損なわれるため，多言語へ拡張するのが容易ではない，(4)二言語辞書の方向性のために，逆方向の翻訳をする場合には辞書をまったく作り直す必要がある，などの欠点を持つ．

[トランスファ方式] 二言語辞書によって両言語の辞書が分離されているために，解析辞書，生成辞書の範囲では言語ごとの語彙知識のモデュラリティが保存されている．したがって，多言語へ拡張は，言語対に依存した二言語辞書だけでできる．しかし，(1)辞書が３種類となり，辞書相互間の整合性の確保が問題となる，(2)解析で分離すべき意味用法と対照辞書による訳語選択で処理すべき意味用法との判断基準が設定できないために，対照辞書の作成には，個々の単語について原言語辞書と相手言語辞書の比較が必要となる．(3)原言語辞書と対照辞書，対照辞書と相手言語辞書の間で記述の重複が生じる，などの欠点がある．

[中間言語方式] 作成・管理すべき辞書がいずれも単一言語辞書となるので，原則的にはその言語についての知識だけで辞書が記述でき，言語ごとの知識のモデュラリティがある．しかしながら，原言語と相手言語の語彙が第３の言語である中間言語の語彙（概念語）を介して結び付けられるが，(1)この第３の言語の語彙はそれ自身具体的な用法も定義も持たないものであるために，原言語の語彙から概念語への写像，および，概念語から相手言語の語彙への写像を規定する辞書記述の部分が，通常の二言語対照辞書の記述以上に困難になる．結局，『言語Ａの単語ａの意味１と単語ａ'の意味２とが表現する概念』ということから，相手言語の生成辞書を作ることになり，単一言語辞書としてのモデュラリティが損なわれる．また，多言語へ拡張すると概念語の定義が『言語Ａの単語ａの意味１と言語Ｂの単語ｂ…の表現する概念』といったものになり，多くの

言語の知識を持った人でないと辞書記述ができなくなる．(3)新たな言語が追加されるごとに概念語の再定義が必要になることが考えられ，この場合には，既存の言語用の辞書すべてを変更しなければならない．結局，個別言語の語の意味をすべての言語に共通な概念語へ写像したことで，この概念語を通じてすべての言語の辞書記述が関係してしまうことになる．

以上のように，各方式ともにその純粋な形では辞書の作成や管理に難点がある．特に現時点では，純粋な中間言語方式をとることはあり得ず，Mu-プロジェクトを含め多くのシステムでは，原言語の単語から相手言語の単語への写像はトランスファ方式を採っている．また，トランスファ方式においても，生成辞書の関連語欄の記述を豊富にして，相手言語内での語彙の言い替えを生成過程で頻繁に行うようにすると，語彙的トランスファ自体は二言語間で行われるが，中間言語方式に近くなる．ただ，辞書作成と管理の点から見ると，定義の曖昧な記号系を設定するたびに作成・管理の負担が増大する．Mu における深層格や意味マーカもその例であるが，これが概念語という数万のオーダの新たな記号系が入ってくると，原言語・相手言語・中間言語という3つの言語の語彙を管理しなければならなくなり，その負担は極めて大きくなろう[20]．

5. おわりに

システランの EC での評価では，誤り原因の 50% が辞書にあったと報告されている[21]．辞書が原因の誤りは，その単語がある特定の環境で使われない限り顕在化しないことから，システム運用後も頻繁に生じ得る．また，システムの新たな分野への適用や改良は単語の新規登録や辞書記述の見直しを必要とすることから，辞書作成のコストや管理の容易さは実用上非常に大きな問題となる．このような辞書の作成や管理の問題を解決するためには，次のような研究を行ってゆく必要がある．

(1) 辞書作成・管理支援システム：大規模辞書を作成，管理してゆくためのソフトウェアは，知識工学における知識獲得と共通の，しかも，それよりも複雑な多くの問題を解決しなければならない．また，実用的なMTには，ユーザが簡単に語彙を登録できるユーザ用の辞書作成支援システムが不可欠である[22),23),24)]．

(2) 記述基準の明確化：記述の基準の問題は，辞書を作成・管理してゆく際の最大の問題である．それぞれ1つ1つを言語学者との共同研究を通じて解決してゆく必要がある[25)]．

(3) 記述の分野依存性：専門用語といった問題だけでなく，単語の用法や意味の記述という辞書の内部の記述にまで分野依存性がある．R. Kittredge らの部分言語 (sub-language) の考え方からの研究，複数の辞書へのアクセスを階層的に行うといったソフトウェア的な研究，頻度情報の活用，などさまざまな観点から研究してゆく必要がある．

(4) 翻訳家用の専門用語辞書：ヨーロッパ・カナダにおいては，専門用語の定義や用例，その訳語，などを記述した翻訳家用のターミノロジ・データ・ベースが数多く作成され[26)]，翻訳家に提供されつつある．このようなデータ・ベースと現在のMTとを統合したシステムの開発，あるいは，これらからMT用辞書を自動作成する研究，などが考えられる[27)]．

(5) 検索効率の良い記憶方式：本稿では取上げなかったが，辞書が大規模になるに伴って，辞書データをどのように記憶しておくかは，記憶容量・処理速度を良くする上で，最も重要な因子となる．これまでにも，各単語の共通の pre-fix 部分を共有して記憶しておく Trie 構造[28)]や平衡木 (B-Tree) を日本語辞書向きに拡張した手法[29)]などが提案されているが，専門分野別の辞書，高頻度語・低頻度語の辞書，などの記憶方式と組み合わせて，今後とも積極的な研究を進めてゆく必要がある．

なお，本稿をまとめるにあたっては，Mu-プロジェクトの辞書 WG での議論を参考にした．特に，名詞意味マーカの設定については，石川徹也氏 (図書館情報大)・烏海剛氏 (JICST)，動詞の格枠の議論については，木村睦子氏 (IBS)・坂本義行氏 (ETL) に負うところが大きい．また，実際の作成，管理については，佐藤雅行氏 (JICST)，中村順一氏 (京大) に負うところが大きい．記して，感謝する．

参考文献

1) 辻井潤一：機械翻訳，別冊『ワープロと日本語処理』，bit (1985)．
2) 長尾 真他：科学技術庁機械翻訳プロジェクトの概要，情報処理，本特集号 (1985)．
3) 内田裕士：言語に依存しない概念構造を中間表現の基本とし，常識を使う多言語向き機械翻訳システム，日経エレクトロニクス，12-17 (1984)．
4) Haberman, G. V. et al.: Systran System of Automatic Translation and its Application at KfK, KfK Karlsruhe (1983).
5) 日本科学技術情報センタ，他：日英科学技術用

辞書データベースの開発に関する報告書 (1984.)
6) 岡本哲也他：語彙関数を用いた日本語の言い替え系，情報処理学会，自然言語処理研究会資料，45-3 (1984).
7) 西田富士夫：言語情報処理，コロナ社 (1981).
8) 長尾 真他：機械翻訳における訳語選択と構造変換過程，情報処理 (1985).
9) 野垣出他：動詞意味素性の付加による日本語文アスペクトの解析および翻訳について，情報処理学会，自然言語処理研究会資料，48-4 (1985).
10) 石綿敏夫：日本語の生成語い論的述語と言語処理への応用，国研報告 54 (1975).
11) Alexander, D. et al.: Some Classes of Verbs in English, Linguistic Research Project, Indiana University (1964).
12) Householder, F. et al.: More Classes of Verbs in English, Linguistic Research Project Indiana University (1965).
13) Isabelle, P. et al.: Taum-Aviation, Description, d'un Systeme de Traduction automatisee des Manuels d'entretien en Aeronautique, TAUM, Universite de Montreal (1978).
14) 田中穂積他：基本動詞『MAKE』を含む文の日本語への訳し分け，情報処理学会，自然言語処理研究会資料，43-3 (1984).
15) Nakamura, J et al.: Grammar Writing System (GRADE) of Mu-Machine Translation Project, Proc. of Coling 84, Stanford (1984).
16) 辻井潤一：訳語選択について，情報処理学会，自然言語処理シンポジウム (1983).
17) 山梨正明他：格解釈と認知機構，第2回認知科学会全国大会予稿集 (1985).
18) 村木新次郎他：辞書における格情報の記述，情報処理学会，自然言語処理研究会資料，46-3 (1984).
19) 片桐他：Mu プロジェクトにおける翻訳実験支援環境，情報処理学会，自然言語処理研究会資料，47-9 (1985).
20) 村木一至：機械翻訳用辞書構成，情報処理学会，自然言語処理研究会資料，46-1 (1984).
21) Slype, G. V. et al.: Description du Systeme de Traduction Automatique Systran de la Commision des Communautes Europeens, Bureau Marcel van Dijk.
22) 小暮 潔他：辞書編集用フレームエディタ，情報処理学会，自然言語処理研究会資料，45-1 (1984).
23) 中村順一：機械翻訳のソフトウェア環境，本特集号 (1985).
24) 坂本義行他：Mu プロジェクトにおける総合システムの基本設計，情報処理学会，自然言語処理研究会資料，46-6 (1984).
25) Ishiwata, T. et al.: Basic Specification of the Machine Readable Dictionary TR-100, Technical Report of ICOT (1985).
26) UNESCO: Terminology Manual, UNESCO & INFOTERM, Paris (1984).
27) 鶴丸弘昭他：単語の釈義文を利用した単語間の階層関係の抽出について，情報処理学会，自然言語処理研究会資料，45-4 (1984).
28) Knuth, D. E.: The Art of Computer Programming, Addison-Wesley Publishing Co., Vol. 3 (1973).
29) 日高 健他：拡張 B-tree による日本語単語辞書の作成，情報処理学会，自然言語処理研究会資料，33-8 (1982).

(昭和60年6月18日受付)