

# 日本語重文・複文を対象とした 文法レベル文型パターンの被覆率特性

池原 悟<sup>†</sup>      徳久 雅人<sup>†</sup>      竹内（村本） 奈央<sup>††</sup>      村上 仁一<sup>†</sup>

日本語の重文と複文（但し、述部を2つ又は3つ持つものに限る）に対して文法レベルで記述された22.1万件の文型パターンを対象に、その被覆率特性を調べ、文型パターン方式の問題点と可能性について検討した。機械翻訳において対訳文型パターンを使用した翻訳方式は、文型パターンに適合した入力文に対して品質の良い訳文が生成できるため、古くから着目されてきた。しかし、被覆率を上げるには、大量の文型パターンが必要なこと、また、汎化を行えば、それにつれて意味的な排他性が失われ副作用が増大することが大きな問題であった。そこで、本稿では、単語レベル、句レベル、節レベルの文型パターン辞書（それぞれ12.1万件、8.8万件、1.1万件で、いずれも文法レベルで記述されている）を対象に被覆率などの特性を評価した。評価パラメータとしては、「任意の入力文に対して適合する文型パターンの有無」を表す「再現率」（「文型再現率」、「文型一致率」の2種）と「適合した文型パターンの意味的な正しさ」を表す「意味適合率」（「適合文型意味正解率」、「適合文型正解含有率」の2種）の4種を定め、それらを使用した。その結果によれば、「文型再現率」は、単語レベル、句レベル、節レベルの順に70%、89%、78%で、いずれもかなり高い値を示すが、入力文に対して多数の意味的に不適切な文型パターン（単語レベルで14件、句レベルで165件）が適合してしまい、適合した文型パターンの中に意味的に正しいもの含まれる割合は、単語レベルで21%にとどまっていることなどが分かった。これらの結果に基づき、「再現率」と「意味適合率」を向上させる方法について検討した結果では、「再現率」は、「任意要素」や時制、相、様相の記述法に大きく依存することが分かった。また、「意味適合率」の向上を図るには、変数に対する強力な意味的制約条件の付与や「原文任意要素」の指定基準の見直しなどが必要であることが分かった。

キーワード： 日英機械翻訳、文型パターン、テンプレート翻訳、被覆率、再現率、適合率、重文、複文

## Study of Cover Ratio of Syntactic Sentence Patterns for Japanese Complex Sentences

SATORU IKEHARA<sup>†</sup>      MASATO TOKUHISA<sup>†</sup>      NAO TAKEUCHI(MURAMOTO)<sup>††</sup>  
JIN'ICHI MURAKAMI<sup>†</sup>

---

<sup>†</sup> 鳥取大学工学部 Faculty of Engineering, Tottori University

<sup>††</sup> フリー言語アナリスト Independent Linguistic Analyst

Pattern based MT has drawn attention for long time since it yields good translations for matched sentences. But it has been difficult problem how to build the pattern pair dictionaries which have a huge number of semantically independent patterns to obtain a high cover ratio. This paper experimentally evaluated the cover ratio of the pattern pair dictionary which has recently been developed for Japanese *Complex* and *Compound* sentences and studied possibility of pattern based MT method. This dictionary contains syntactic sentence patterns of *Word Level* (121,000 patterns), *Phrase Level* (88,000 patterns) and *Clause Level* (11,000 patterns) which are generated from 150,000 example sentence pairs for Japanese to English. Evaluation was conducted by using 4 parameters such as “*Sentence Recall Ratio*,” “*Sentence Coincide Ratio*,” “*Semantic Precision Ratio*,” and “*Matched Pattern Precision Ratio*.” The results are as follows. “*Sentence Recall Ratios*” are 70%, 89% and 78% for each of *Word level*, *Phrase Level* and *Clause Level* sentence patterns, and “*Matched Pattern Precision Ratio*” of *Word Level* sentence patterns is 21%. Though “*Matched Pattern Precision Ratio*” was low, it was clarified that there are many ways left to increase the matched patterns.

**KeyWords:** *Japanese to English Machine Translation, Sentence Pattern, Template MT, Cover Ratio, Recall Factor, Precision Factor, Complex Sentence, Compound Sentence*

## 1 はじめに

従来の機械翻訳システムは要素合成法 (Semantic Composition) を基本としているが、言語表現には意味的に非線形なものが多く、表現を分解して行く過程で全体の意味が失われることが問題であった。この問題を解決するには、文構造とその意味を一体的に扱う仕組みが必要だと考えられる。

このような仕組みとしては、既に、古くから文型パターン翻訳 (テンプレート翻訳とも言う) の方法が試みられてきた。この方法は、文型パターンに適合した入力文に対して品質の良い訳文が得られることから、トランスファー方式と併用する形で多くの商用システムで実現されている。最近では、翻訳メモリとも併用する形式で採用されている例が多い。

しかし、使用されている文型パターンの数はいずれも 100~300 パターン程度で少なく、補助的な仕組みとして使用されるか、もしくは、ある特定の狭い分野の文書に適用されているのが普通である。これは文型パターンの作成コストが高いこともあるが、パターン数が増大するとパターン間の相互干渉が増加して翻訳精度が低下することが主な原因と見られる。

これに対して、文構造と意味の関係の体系化を試みたものとして多段翻訳方式 (池原, 宮崎, 白井, 林 1987) の研究がある。この方式では、動詞または形容詞と共起する名詞の関係を 1.7 万件の結合値パターン (池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 1997) としてまとめており、単文レベルでの訳文品質は従来に比べて大幅に向上した (金出地, 池原, 村上 2001)。しかし、複文 (埋め込み節を持つ文) や重文 (接続構造を持つ文) に対して、意味の非線形性<sup>1</sup>を扱う仕組み

<sup>1</sup> 慣用表現の例からも分かるように、言語表現の意味が、それを構成する要素の意味の和として表現できないこと

みがないこと (Ikehara 2001), また, 意味の単位とされる原言語の表現構造に対して, 単一の目的言語表現が対応づけられるため, 文脈上, 不適切な表現への翻訳が防止できないことが問題として残されている (池原 2001) .

これらの2つの問題を解決することを狙って, 最近, 「等価的類推思考の原理による機械翻訳方式」(池原 2002)(池原 2003) が提案された. この方式は, 原言語と目的言語の表現から非線形な表現構造 (池原 2004) を取り出して文型パターン化した後, 意味の同等性に着目して類型化 (有田 1987) すること, また意味的に類型化された文型パターン間を「類推思考の原理」(市川 1963) によって対応づけることの2つの仕組みから構成される.

この方式を実現するには, 原言語, 目的言語の表現に対する「意味類型パターン」を収録した大規模な「意味類型知識ベース」を構築することが必要である. 「意味類型パターン」は, 原言語表現の意味的に非線形な構造を取り出して, それを意味的に分類体系化し, 原言語と目的言語に共通する概念(「真理項」と呼ぶ)を介して目的言語の「意味類型パターン」に対応づけるものである. 従って, 「意味類型知識ベース」を開発するには, あらかじめ, 原言語と目的言語の対訳文から意味的に非線形な構造を取り出して「対訳文型パターン辞書」を作成することが必要となる.

ところで, このような文型パターン辞書構築の最大の問題は, 「文型パターンの網羅性」と「意味的な排他性」をいかにして実現するかの2点である.

このうち網羅性を実現するには, 汎化された大量の文型パターンが必要と考えられるが, 汎化が進むにつれて排他性を確保することが難しくなることが予想される. このような網羅性と排他性の問題を解決するには, 実際にある程度大量の文型パターンを作成しながら, 実験的に汎化の可能性を調べていくことが必要と考えられる.

これに対して, 新翻訳方式に関する上記の研究では, 第1段階として, 日英対訳例文 100 万件の中から重文, 複文の例文 12.9 万件を抽出し, それを文法的なレベルで汎化することによって単語レベル, 句レベル, 節レベルの文型パターン 24.6 万件(そのうち異なり文型パターン数は 22.1 万件) が試作された (池原, 阿部, 徳久, 村上 2004). そこで, 本研究では, 日本語試験文と文型パターンとの照合実験を行って, 文型パターンの被覆率と排他性について評価し, 文法レベルで定義された文型パターンの問題点と今後の対策について検討する.

以下, 第2章では, 作成された文型パターン辞書の概要を示し, 第3章で被覆率特性の評価方法を述べる. 第4章では, 再現率の評価結果を示し, 第5章で文型パターンを記述する方法と被覆率の関係についての評価結果を示す. 第6章で意味的な適合率について評価する. 最後に第7章で評価結果をまとめ, 正解率を向上させるための方法について考察する.

---

を言う. 言語表現の意味の線形性, 非線形性の定義とその判定の方法は, (池原 2004) で提案されている.

## 2 文型パターン辞書の概要

本章では、文型パターン作成の元となった日英対訳文の内容と作成された文型パターン辞書の概要を示す。

### 2.1 元となった日英対訳標本文

本稿で評価対象とする辞書は、2カ所又は3カ所の述部を持つ<sup>2</sup>日本語の重文（文接続のある文）、複文（埋め込み文のある文）の非線形な表現構造<sup>3</sup>を文法レベルでパターン化したもの<sup>4</sup>で、以下の手順で作成されている。すなわち、日本語の基本的な表現が収録されていると見られる辞書や語学教育用の教科書、機械翻訳機能評価用の試験文等、約30種類のドキュメント（日英対訳文100万件）から該当する範囲の重文と複文の対訳標本文12.9万件<sup>5</sup>を取り出し、それを汎化することによって作成されており、単語レベル、句レベル、節レベルの3種類の文型パターン（異なり文型パターン22.1万件）が収録されている。

文型パターン辞書の作成に使用された標本文の種類と例を表1に示す。

日本語標本文当たりの平均文字数は、23.3字/文（最大148字/文）であり、平均形態素数は、12.9個/文（最大63個/文）である。また、それと対応する英語訳文の平均単語数は、10.3語/文（最大59語/文）である。

### 2.2 文型パターンの記述レベルと汎化の方法

文型パターンは、対訳標本文中の線形な要素を対象に、変数化、関数化、任意化などを行ったもので、変数化された要素の文法的属性に着目して、単語レベル、句レベル、節レベルの3グループから構成される。表2に各レベルにおける主な汎化の内容を示す<sup>6</sup>。

これらの3段階は汎化の程度を表すもので、いずれも使用される変数や関数は、文法的属性によって変域が指定されるものであることから文法レベルで定義された文型パターンである<sup>7</sup>。

<sup>2</sup> 多数の述部を持つ重文、複文の場合、全体が非線形である場合は少なく、多くの場合は、少数の述部を持つ非線形構造の組み合わせによって構成されていると考えられるため、述部2又は3までの非線形構造を対象に文型パターン化されている。

<sup>3</sup> 対象とする文型パターン辞書では、「言語表現中、他の要素に置き換えても表現構造全体の意味が変わらない要素（例えば変数化したとき変域が定義できる要素）」を「線形要素」と定義し、「線形要素のみから構成される表現構造」を「線形な表現構造」と定義している。また、日本語表現の意味を英語表現を用いて定義することにより、具体的な判定基準を設けている。詳細は、(池原 2004) を参照。

<sup>4</sup> 形態素解析で得られる品詞情報と句や節の単位に関する情報を使用して定義されたもので、変数に対する意味的な制約条件などは使用されていない。

<sup>5</sup> 日本語は重文、複文であるが、対応する英文は、重文、複文とは限らない。意識されている例が多いため、英文は単文である場合も多い。

<sup>6</sup> 各レベルの詳細は、(池原他 2004) を参照。また、文型パターン記述方法については、(池原, 宮崎, 佐良木, 池田, 白井, 村上, 徳久 2003) を参照。

<sup>7</sup> 文型要素が形態素解析で得られた文法情報の範囲で記述されている点で「文法レベルの文型」と呼ぶ。これに対して、変数の意味的な制約条件などを付与した文型を「意味レベルの文型」と呼ぶ。これらはいずれも意味的な類型化は行われていないため、「意味類型パターン」とは区別される。

表 1: 対象とした標本文の種類

文種別	述部数	説明	例文
文種別 1	2	文接続 1 カ所を持つ文	・私はいすに座り、窓の外を眺めた。 ・彼の声は大きいので、部屋中に響いた。
文種別 2	3	文接続 2 カ所を持つ文	・明日は東京に行って 友達に会い 久しぶりに話をする。 ・電話をしながらメモを取ったが そのメモをなくしてしまった。
文種別 3	2	埋込み文 1 つを持つ文	・さっきまでここにいた猫がいなくなった。 ・お客様にお渡しする品物は、丁寧に扱って下さい。
文種別 4	3	埋込み文 2 つを持つ文	・約 20 名を乗せた飛行機が、人の住んでいない島に不時着した。 ・父と撮った写真をアメリカにいる姉に送った。
文種別 5	3	文接続, 埋込文各 1 を持つ文	・彼女の吐き出す息は白く、頬は赤かった。 ・母が作ってくれたマフラーを見ると 学生時代を思い出す。
< 凡例 > _____ , _____ : 文接続の従属節 . [ ] : 埋め込み文の従属節			

### 2.3 評価対象とする文型パターン数と例

前項の方法によって作成された文型パターン数の内訳を表 3 に示す。また、各レベルの文型パターンの例を表 4 に示す。単語レベル、句レベル、節レベルと汎化が進むにつれて文型パターンは縮退するため文型パターン数は減少している。特に節レベルの文型パターン数は少ないが、これは節レベルまで汎化できるような標本文が少なかったためである<sup>8</sup>。

## 3 調査実験の方法

### 3.1 実験の方法

本実験では、「文型照合プログラム<sup>9</sup>」を使用して、日本語重文、複文を文型パターン辞書と照合し、適合する文型パターンを抽出する。その結果から文型パターン辞書の被覆率特性を求める。

テスト用の入力文としては、文型パターンの作成に使用した対訳標本の日本語（以下では「母集団試験文」と言う）を使用する。但し、この場合、文型パターンは対訳標本文を汎化することによって作成されているから、各文型パターンはその元となった標本文と必ず適合する。

<sup>8</sup> 意識によって単文で翻訳されている標本文が多い。そのような標本文では、日英双方で節間の意味的な対応関係が決めることが困難で、節レベルの文型パターン化はできない。

<sup>9</sup> 文型パターン辞書の中から入力文に適合する文型パターンをすべて発見して抽出するプログラムである。入力文が与えられたとき、指定された要素のすべてが、指定された順に出現する文型パターンがその入力文に適合した文型パターンと判定される。

表 2: 文型パターンの記述レベルと汎化の内容

レベル	汎化の主な内容
単語 レベル	線形要素の自立語を変数化した文型パターンで、おおよそ以下の汎化を行ったもの。 (1) 名詞, 動詞, 形容詞, 副詞などの自立語を変数化 (2) 線形で文型上不要な要素を任意化し, 文型の骨組みとなる要素を抽出 (3) 字面要素について可能な範囲でグループ化
句 レベル	句の変数は使用されているが, 節の変数の使用されていない文型パターンで, 単語レベルの文型パターンにおおよそ以下の汎化を行ったもの。 (1) 適用範囲を品詞から句へ拡大 (品詞変数は可能な範囲で句変数化) (2) 機能語の適用拡大 (格助詞 格助詞相当語, 等) (3) 英語句生成関数の適用
節 レベル	節の変数が一つ以上使用された文型パターンで, 句レベルの文型パターンにおおよそ以下の汎化を行ったもの。 (1) 名詞節, 副詞節, 主節, 従属節の 1 変数化による重文, 複文の基本構造をパターン化 (2) 日本語節から英語句への変換関数の使用 (3) その他英語構造生成関数の使用

そこで, 評価では, テストに使用する入力文から作成された文型パターンを除いた文型パターン辞書を使用する<sup>10</sup>。

<sup>10</sup> 具体的には, 該当する入力文から作成された文型パターン (「自己パターン」と言う) への適合はカウントせず, それ以外の文型パターンへの適合のみをカウントする。但し, 入力試験文とは異なる標本文から作成された文型パターンに「自己パターン」と同一のものであればそれは「自己パターン」とは見なさない。この方法は, クロスバリデーションの方法で, 同種の母集団を対象としたオープンテストである。

表 3: 対訳標本文数と作成された異なり文型パターン数

	日英対訳 標本文数	作成された異なり文型パターン数			
		単語レベル	句レベル	節レベル	合計
文種別 1	57,235 (44%)	53,578 (44.0%)	37,356 (42.3%)	5,521 (48.1%)	96,455 (43.7%)
文種別 2	6,196 (5%)	6,080 (5.0%)	4,952 (5.6%)	417 (3.6%)	11,450 (5.2%)
文種別 3	46,907 (36%)	44,008 (36.2%)	30,932 (35.0%)	3,185 (27.7%)	78,125 (35.3%)
文種別 4	5,986 (5%)	5,889 (4.8%)	5,084 (5.8%)	811 (7.1%)	11,784 (5.3%)
文種別 5	12,389 (10%)	12,174 (10.0%)	10,025 (11.3%)	1,551 (13.5%)	23,750 (10.7%)
- -	128,713 (100%)	121,729 (100%)	88,349 (100%)	11,485 (100%)	221,563 (100%)

### 3.2 評価基準

#### 評価パラメータの種類

入力文と文型パターンの照合実験では、入力文に適合した文型パターンが、必ずしも意味的に正しいとは言えない。例えば、以下の例では、入力文に対してそれぞれ2つの文型パターンが適合するが、いずれの場合も正しいのは<適合文型1>であり、<適合文型2>は正しくないため、対応する英語文型パターンは翻訳に使用できない。

表 4: 日英文型パターンと適合する文の例

区 別		日本語文型パターン又は日本文	英語文型パターン又は英文
単 語	文型パターン	それは/ <i>N1</i> に/あるまじき/ <i>N2.da</i>	Such <i>N2</i> be unseemly for <i>N1</i> .
	言語表現例	それは学生にあるまじき行為だ。	Such behavior is unseemly for students.
レ ベ ル	文型パターン	#1[ <i>N1</i> は]/ <i>N2</i> の/ <i>V3</i> /ことも/ #2[ある/程度は]/ <i>V5</i> .	#1( <i>N1</i>  I) can <i>V5</i> what <i>N2</i> <i>V3</i> #2[to some extent].
	言語表現例	次郎の言うこともある程度はわかる。	I can understand what Jiro says to some extent.
句 レ ベ ル	文型パターン	あれこれ/ <i>V1.temiru.kako</i> が/ <i>N2</i> が/ <i>NP3.da</i>	All things <i>V1.past</i> , <i>N2</i> 's <i>NP3</i> .
	言語表現例	あれこれ考えてみたがそれがいちばん いい解決策だ。	All things considered, that's the best solution.
	文型パターン	#1[ <i>N1</i> は]/ <i>NP2</i> も/ <i>V3.nai</i> とは/ <i>VP4.gimu</i> /ことだ。	It is <i>AJP</i> ( <i>VP4</i> ) that #1( <i>N1</i>  you) should <i>V3.not</i> <i>NP2</i> .
言語表現例	総理大臣の名前も知らないとは まことに哀れむべきことだ。	It is really pitiable that you should not know the Prime Minister's name.	
節 レ ベ ル	文型パターン	#1[ <i>N1</i> は]/ <i>CL2</i> とは/ <i>V3.nai.kako</i>	#1( <i>N1</i>  I)did <i>V3.not</i> <i>CL2.past</i> .
	適用例	彼があればほど英語が話せるとは 思わなかった。	I didn't know he could speak English so well.
節 レ ベ ル	文型パターン	<i>CL1.teiru.nai.da</i> と/ <i>N2</i> は/ <i>VP3.kako</i>	<i>N2</i> <i>VP3.past</i> that <i>CL1.not</i> .
	言語表現例	彼女はもうぼくを愛していないのだと ぼくは自分に言い聞かせた。	I convinced myself that she did not love me any more.
<p>&lt;凡例&gt; <math>N_n</math> : 名詞変数, <math>V_n</math> : 動詞変数, <math>NP_n</math> : 名詞句変数, <math>VP_n</math> : 動詞句変数, <math>CL_n</math> : 節変数,  <math>.abc</math> : 時制・相・様相および語形変形の関数, <math>\hat{abc}</math> : 活用形指定関数,  <math>\#1(N_n   I)</math> : 「選択的訳出記号」で, <math>N_n</math> の値がないときは, 訳語として "I" を使うことを指示.  <math>\#n[ ]</math> : 「文型任意要素記号」で, なくても良い要素を指定する記号で,  主語, 目的語の補完でも使用される,  <math>/</math> : 「原文任意要素記号」で, この位置に任意の表現の挿入が認められる.</p>			

例 1 ) 係り受け構造の異なる複数の文型パターンに一致する例

入力文 : 「キャッチャーはバットが回ったとアピールした。」

<適合文型 1 > :  $N1$  は/ $N2$  が/ $V3.kako$  と/ $V4.kako$ .

$N1$   $V4.past$ , saying  $N2$  had  $V3.past$ .

The catcher appealed, saying the batter had swung.

<適合文型 2 > :  $\#6[N1$  は] $N2$  は/ $N1$  が/ $V3.kako$  と/ $V4.kako$ .

( $N1$ |I)  $V4.past$  that  $N1$   $V3.pft$   $N2$ .



The batter appealed that the batter had swung the catcher. (不適切)

(文型2の適例) それは自分がやったと打ち明けた。 I confied that I had done it.

【説明】 入力文の「は格」は、動作主として、文末の述部「アピールする」に掛かる。＜適合文型1＞は、これと同じ係り受け構造をもち、対応する英語文型パターンから、例にあるような訳文が生成される。しかし、＜適合文型2＞の「は格」(N2は)は、V4の目的格であり、V4に掛かる。このため、入力文の構造と異なり、対応する英語文型パターンは使用できない、なお、＜適合文型2＞の#6[N1は]は、省略補完の役割を持ち、入力文に主格の名詞があるときは、それをN1にバインドすることを意味している。

例2) 主節の述部の意味によって訳し分けるべき複数の文型パターンに適合する例

入力文: 「自分でその責任をとることは明らかだ。」

<適合文型1>: N1でREN2 N3を/V4ことは/明らかだ。

It is clear to V4 AJ2 N3 by N1.reflex.

It is clear to take that responsibility by myself.

<適合文型2>: N1は VP2ことは/AJV3。

(N1|I) V(AJV3) VP2.ing.

I clear taking that responsibility by myself. (不適切)

(文型2の適例) ここで働くことは好きだ。 I like working here.

<凡例> REN<sub>n</sub>: 連体詞変数, AJV<sub>n</sub>: 形容動詞変数, .reflex: 再帰表現変換関数,  
: ゼロ代名詞などの補完主語を表す記号

【説明】 <適合文型1>は、文末の述部「明らかだ」が字面で定義された文型パターンに一致しており、対応する英語表現 “It is clear that ~” を使用した翻訳ができる。この入力文は、文末述部が変数化された<適合文型2>にも適合するが、この文型は主節の述部を動詞に変換して訳するため、入力文の翻訳に適さない。

そこで、被覆率を「再現率」と「意味適合率」に分けて評価する。但し、「再現率」は、「入力文の何パーセントに対して適合文型パターンが存在するか」を表し、「意味適合率」は、「適合文型パターンのうちどれだけの文型パターンが意味的に正しいか」を表す<sup>11</sup>。

このうち、まず「再現率」の評価方法であるが、文型照合プログラムは、入力文と文型パターンとの照合において、当該文型パターンのすべての要素が指定された順に出現するような入力文である時、その文型パターンを適合文型パターンと判定する。すなわち、ある文型パターンが入力文との照合に成功するには、文型パターンで指定された要素がすべて入力文に存在しなければならないが、逆に、入力文のすべての要素が文型パターンで指定されている必要はない。そこで、適合文型パターンを以下の2種類に分類する。

<sup>11</sup> 「再現率」は見かけのカバー率を表す意味で、情報検索で 사용되는「再現率」の意味とは異なる点に注意。これに対して「意味適合率」は、情報検索の「適合率」の意味と同様である。なお、真の意味でのカバー率は「再現率」×「意味適合率」で評価される。

< 「完全一致」する文型パターン >

入力文のすべての要素が文型パターンの要素と適合する文型パターン。但し、文型パターン中の「文型任意要素」に該当する要素は、入力文中になくても良い。

< 「部分一致」する文型パターン >

入力文の一部の要素が指定されない文型パターン。すなわち、入力文が、「原文任意記号（離散記号）」の部分に該当する部分を持つ文型パターン。

下記に入力文に「完全一致」する文型パターンと「部分一致」する文型パターンの例を示す。

例) 入力文に「完全一致」又は「部分一致」する単語レベル文型パターンの例

入力日本語：「その飛行機は荷物を積み残して出発した。」

< 完全一致する文型パターン >

$W_n$  : は単語レベルの文型パターン,  $P_n$  は句レベルの文型パターンを意味する。

(W1)  $REN1/ N2$  は/  $N3$  を/ 積み残して/  $V4.kako$ 。

その 飛行機は 荷物を 積み残して 出発した。

$AJ1 N2 V4.past$  before some  $N3$  could board.

(W2)  $\#1[REN2]/ N3$  は/  $\#4[NUM5$  の]/  $N6$  を/  $V7$  て/  $V8.kako$ 。

その 飛行機は 荷物を 積み残して 出発した。

$\#1[AJ2] N3 be.past V8.ed$  by  $V7.ing \#4[NUM5] N6$ .

(P1)  $NP1$  は/  $N2$  を/  $\#3[大きく]/ V4$  て/  $VP5.kako$ 。

その飛行機は 荷物を 積み残して 出発した。

With  $N (V4)$  of  $NP1.poss N2 NP1.pron VP5.past$ .

(P2) <  $N1$  は > /  $VP2$  て/  $VP3.kako$ 。

その飛行機は 荷物を積み残して 出発した。

$(N1 | I) VP2.past$  and  $VP3.past$ .

< 凡例 >  $NUM_n$  : 数詞変数,  $N (V_n)$  : 変数を引数とする「変数関数」で、引数の変数値（この場合は動詞）を変数名の品詞（この場合は名詞）に変換する。

【説明】いずれも入力文の全要素が文型パターンの要素に一致する。但し、W2, P1 では  $\#n [ ]$  の部分に該当する入力文要素はないが、これらは「文型任意要素」なので、なくとも良いと判断される。

< 部分一致する文型パターン >

文型パターンに存在しない入力文の要素を《 》の記号で示す。

(W1)  $REN1/ N2$  は /  $V3$  て/  $V4.kako$ 。

その 飛行機は《荷物を》積み残して 出発した。

$AJ1 N2 be .past V3.ed$  and  $V4$ .

(W2)  $\#1[REN2]/ N3$  は/  $\#4[N5$  の]/ $\#6[AJV7]/ N8$  を /  $V9.kako$ 。

その 飛行機は 荷物を《積み残して》出発した。

$\#1[AJ2] N3 V9.past \#6[AJ7] N8 \#4[of the N5]$ .

(P1) NP1 は/ VP2 / V3.kako。  
その飛行機は 荷物を積み残し《て》出発した。

NP1 VP2.past and V3.past.

(P2) NP1 / VP2 て/ VP3.kako。  
その飛行機《は》荷物を積み残して 出発した。

NP1 VP2.past for VP3.ing.

【説明】《 》の記号で表された要素は、「離散記号“/”」と照合され、「原文任意要素」と判定される。また、W2の“#4[N5の]/#6[AJV7]”の部分は省略できる要素なので、不足する要素とは見なされない。文型パターンによる入力文のカバー率は（入文文字数・《 》部分の文字数）/（入力文文字数）で計算される。

このように、適合文型パターンだとは言え、必ずしも入力文のすべての要素に対して解釈を与えるものではないから、適合文型パターンについて、それが入力文の何%をカバーしているか（入力文カバー率）が問題となる<sup>12</sup>。そこで本稿では、「再現率」を、「入力文単位で見た再現率」を意味する「文型再現率」と「入力文の文字単位で見た再現率」を意味する「文型一致率」の2つのパラメータを使用して評価する。

次に、「意味適合率」であるが、適合した文型パターンには、すでに述べたように、対応する英語文型パターンが翻訳に使用できるものと、そうでないものが存在するため、翻訳に使用できる文型パターンを選択する方法が問題となる。そこで、ランダムに選択する場合（最悪の場合）ともっとも適切なものが選択できた場合（最良の場合）を考え、評価では、「適合した文型パターンの一つ一つが意味的に正しい<sup>13</sup>かどうか」を表す「適合文型意味正解率」と「1入力文に適合した文型パターン（1つ以上）の中に意味的に正しい文型パターンが1つ以上含まれる割合」を示す「適合文型正解含有率」を評価パラメータとして使用する。

以下、「再現率」と「意味適合率」を評価するためのパラメータの計算方法を示す。

#### 被覆率評価のためのパラメータ

<文型再現率：R1>（文単位に見た再現率）

入力文に対して適合文型パターンが存在するかどうかを文単位で集計したもので、下式で定義される。

$$\text{文型再現率 } R1 = M/N$$

N：テスト用入力文の数

M：「自己パターン」以外に適合する文型パターンがあった入力文の数

<sup>12</sup> 文型パターンにカバーされていない要素は、別途翻訳して英文に組み込むことが必要であり、その場合は、どこにどのように組み込めばよいかを判定するため、構文解析情報などが必要となる。

<sup>13</sup> 「意味的に正しい」（又は「意味的に適合する」）文型パターンといえば、対応する英語文型パターンが入力文の英訳に使用できるような文型パターンのことである。

< 文型一致率 :  $R2$  > ( 文字単位に見た再現率 )

「入力文の文字単位に見た再現率」を表すもの<sup>14</sup>で、以下の式で定義される。

$$\text{文型一致率 } R2 = \frac{\sum_{j=1}^M M_j}{\sum_{i=1}^N L_i}$$

$N, M$  :  $R1$  の場合と同じ

$L_i$  :  $i$  番目の入力文の文字数

$M_j$  : 「最尤文型パターン」と一致する入力文の文字数

但し、「最尤文型パターン」とは、「自己パターン」以外の適合文型パターン（一般に複数）のうち、最も広範囲（文字数で計算）に入力文と一致する文型パターンを言う。なお、上記の  $M_j$ 、 $N_i$  の値としては、入力文中の単語数を使用することも考えられるが、ここでは、簡単のため、文字数を使用することにした<sup>15</sup>。

< 適合文型意味正解率 :  $P1$  >

入力文に適合した「自己パターン」以外の任意の文型パターンが正しい文型パターンである確率で、下式で定義される。

$$\text{適合文型意味正解率 } P1 = \frac{1}{M} \sum_{i=1}^M \frac{n_i}{N_i}$$

$M$  :  $R1$  の場合と同じ

$N_i$  :  $i$  番目の入力文に適合した文型パターンの数 ( $N_i \geq 1$ )

$n_i$  :  $i$  番目の文に適合した文型パターンのうちの正しい文型パターンの数 ( $\geq 0$ )

< 適合文型正解含有率 :  $P2$  >

適合した文型パターンの中に、正しい文型パターンが1つ以上含まれる確率で、以下の式で定義される。

$$\text{適合文型正解含有率 } P2 = m/M$$

$M$  :  $R1$  の場合と同じ

$m$  : 正しい文型パターンが一つ以上存在した入力文の数

<sup>14</sup> 従来機械翻訳では、文単位に翻訳率を計算する場合が多い。しかし、通常、長文の翻訳正解率は悪いのに対して単文の正解率は良いため、この方法では、正解率は実際以上に高く評価されることが問題である。「文型一致率」は、この点に留意したもので、再現率を単語数又は文字数で計算している。

<sup>15</sup> テスト文として 12.9 万件を使用した実験結果によれば、形態素単位の評価と文字単位の評価の結果は極めて良く一致し有意差は見られなかったので、文型照合アルゴリズム上計算の単純な文字単位での評価とした。

## 4 日本語文型パターンの被覆率特性

本章では、「文型再現率  $R1$ 」と「文型一致率  $R2$ 」を評価する。実験では、文型パターンの作成に使用した日英対訳文 12.9 万件から、「文種別毎の文数の割合を保つこと」、「それぞれの文種別内での文当たりの文字数の分布が保たれること」を条件に、合計 1 万件の日本文を抽出し、テスト用の入力文として使用した。

### 4.1 被覆率特性と飽和特性

#### (1) 文型パターンの被覆率

前述の 1 万件の日本語入力文に対して単語レベル、句レベル、節レベルの文型パターンの「文型再現率  $R1$ 」、「文型一致率  $R2$ 」を求めた。その結果を表 5 に示す。

表 5: 文型パターンの被覆率

レベル	文型再現率 $R1$	文型一致率 $R2$
単語レベル	69.8 %	47.2 %
句レベル	89.0 %	76.2 %
節レベル	78.1 %	63.1 %
混合レベル	91.8 %	78.9 %

表 5 で「混合レベル」は、単語レベル、句レベル、節レベルの文型パターン全体（異なり 22.1 万件）に対する照合実験の結果を示す。この表から以下のことが分かる。

- (1) 単語レベル、句レベルの「文型再現率  $R1$ 」は、約 70%、89%で高い値を示す。
- (2) これに比べて最も汎化されたはずの節レベルの「文型再現率  $R1$ 」78%は比較的小さい。
- (3) また、「文型一致率  $R2$ 」は、それらより 13~23%低下する。
- (4) 「混合レベル」の「文型再現率  $R1$ 」、「文型一致率  $R2$ 」は、いずれも句レベルの場合より 2~3%向上しただけである。

以上、文数から見て、入力文の 7 割が単語レベルの文型パターンに適合し、9 割が句レベルの文型パターンに適合するのに対して、もっとも汎化されている節レベルの文型パターンの被覆率はその割に低い。節レベルの文型パターンは、被覆率の向上を狙って作成されたものであるが、実際の対訳例文において節レベルまで汎化できるものが少なかったため、結果的に句レベルよりも小さい被覆率しか得られなかったものと考えられる。

また、混合レベルの被覆率が句レベルに比べて余り向上していないが、これは、単語レベ

ルで適合する文の大半が句レベルでも適合すること、また、句レベルの文型パターンに比べて節レベルの文型パターンの数が少なく、句レベルの文型パターン以上に広い範囲をカバーできていないことを意味する。なお、混合レベルでの被覆率が90%を超えていることから、文型パターン数と被覆率の関係はほぼ飽和状態に達しており、これ以上標本量を増やしても被覆率は簡単には向上しないのではないと思われる。

以下、文型パターンの汎用度と文型パターンに定義されない要素について考察する。

#### a) 各レベルの文型パターンの汎用度

上記(2)で述べたように節レベルの文型パターンの被覆率は比較的小さいが、1文型パターンあたりの汎用性はかなり高いと思われる。そこで、文型パターンの「汎用度」を「文型再現率  $R1$  / 文型パターン数」に比例するとし、単語レベルの「汎用度」を1とすると、句レベル、節レベルの「汎用度」は、それぞれ1.8, 12.5となる。すなわち、句レベルの文型パターンは単語レベルの文型パターンより約2倍の汎用性があり、節レベルの文型パターンはさらにその6倍程度の汎用性があることになる。

#### b) 文型パターンにカバーされない入力文要素

既に述べたように、適合した文型パターンによって入力文のすべての要素が解釈される訳ではない。「文型再現率  $R1$ 」と「文型一致率  $R2$ 」の差から計算すると、平均して、単語レベルでは32%(=1 - 47.2%/69.8%)、句レベルでは14%(=1 - 76.2%/89.0)、節レベルでは、19%(=1 - 63.1%/78.1%)の文要素が文型パターンでは解釈されない要素として残されることになる。

### (2) 被覆率の飽和特性

次に、「異なり文型パターン数」と「文型一致率  $R2$ 」(×印)の関係を図1に示す。図中、文型パターン数2.5万件以下の部分(波線)は外挿したものである。また、参考のため、全体の文型パターンでの「文型再現率  $R1$ 」(○印)の値も図示した。

これより、以下のことが分かる。

- (1) 単語レベル、句レベルのいずれの文型パターンも、「異なり文型パターン数」が数万件になると飽和傾向が現れる。
- (2) いずれの場合も「異なり文型パターン数」が1万件以下では、有効な被覆率は得られそうにない。
- (3) 「文型一致率  $R2$ 」は、まだまだ向上の余地は残されているが、現在の文型パターン化の方法では、現状(10万件程度)の文型パターン数をこれ以上増加させても、被覆率の向上はあまり期待できそうにない。

以上から、これ以上標本文数を増加させても、その割に被覆率は向上しないとみられる。前項で述べたように、むしろ、汎化方法について、さらに改良の方法を検討することが重要と判断される。

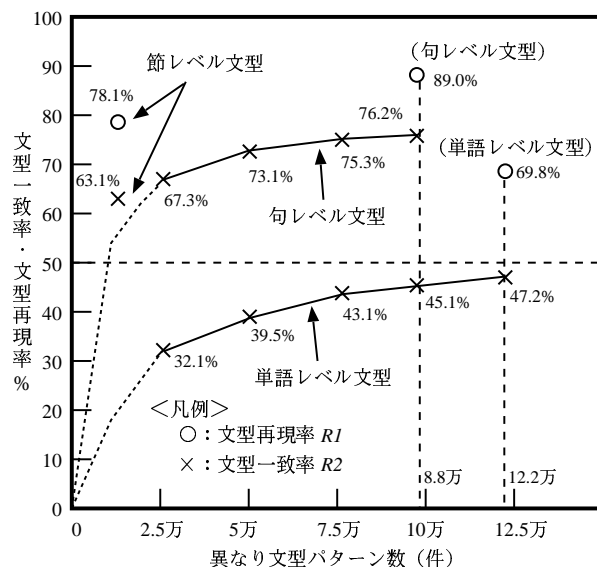


図 1: 文型一致率 (文字数で見た再現率)

### (3) 文型パターン種別と被覆率特性の関係

表 5 の結果について、その内訳を調べるため、単文レベルの文型パターンについて文種毎の「文型再現率  $R1$ 」を求め、その結果を表 6 に示す。

表 6: 文種別毎に見た単語レベルの文型再現率  $R1$

文種別	1	2	3	4	5	全体
文型再現率 $R1$	70.1%	63.2%	71.8%	65.2%	65.8%	69.8%

この表から、単語レベル、句レベル共に「文型再現率  $R1$ 」は文種別に依存せず、ほぼ一定であり、バランスの良い標本集合となっていることが分かる。ところで、直感的に言えば、文種別 2, 4, 5 は、述部を 3 つ持つ重文、複文を対象としたものであるから、表現は、述部数 2 の文種別 1, 2 に比べてさらに複雑で、多様だと予想される。しかしながら、これらの文種別で作成された文型パターン数は、文種別 1, 3 の場合の  $1/5 \sim 1/10$  程度となっており (表 5)、逆に、かなり少ない。このように、より多様と思われる述部 3 以上の表現が、述部 2 の場合よりも少ない文型パターンで同じ程度カバーされることから、構文構造を考えたとき、日本語表現では、述部数が増えてもその割に表現構造の多様性が増加するわけではないと推定される。これは、文型パターン方式の可能性を期待させるものと言うことができる<sup>16</sup>。

<sup>16</sup> 文型パターンは、非線形な文型構造の持つ意味をすくい取る網のようなものである。もし、述部の多い長文にな

## 4.2 適合の仕方と適合文型パターン数

### (1) 「完全一致」と「部分一致」の割合

すでに、3.2節で述べたように、入力文と文型パターン辞書との照合では、適合した文型パターンとして「完全一致」した文型パターンと「部分一致」した文型パターンが抽出される。前者の場合、入力文の全要素が適合した文型パターンを使用して翻訳されるが、後者の場合は、入力文要素の中に文型パターンに定義されない要素が含まれているため、それらの要素は文型パターンとは別の方法で翻訳して目的言語に組み込まなければならない。そのため、入力文に対していずれの形で適合したかが問題となる。

そこで、単語レベルの文型パターンとの照合において、入力日本文に対して、「完全一致」する文型パターンが存在した割合と「部分一致」する文型パターンしか存在しなかった割合を調べた。その結果を表7に示す。

表7: 「完全一致」と「部分一致」の割合

レベル	完全一致	部分一致	合計 (文型再現率 $R1$ )
単語レベル	15.0 %	54.8 %	69.8 %
句レベル	54.3 %	34.6 %	89.0 %
節レベル	39.5 %	38.6%	78.1 %
混合レベル	56.2 %	35.6 %	91.8 %

表中の「完全一致」の欄は、入力文の中で、「完全一致」した文型パターンがひとつ以上あった文の割合を示し、「部分一致」の欄は、「部分一致」した文型パターンしか存在しない入力文の割合を示す。「文型再現率  $R1$ 」は、「完全一致」した文型パターン又は「部分一致」した文型パターンのいずれかを持つ入力文の割合であるので、両者の合計に一致する。

この表から以下のことが分かる。

- (1) 単語レベルの文型パターンでは、「完全一致」に比べて「部分一致」となる割合は4倍近い。
- (2) これに対して、句レベルの文型パターンでは、この関係が逆転しており、「完全一致」の割合は1.5倍である。
- (3) 節レベルの文型パターンでは、「完全一致」と「部分一致」の割合が拮抗している。
- (4) また、混合レベルは句レベルに近い値である。

これらは、各文型パターンの性質をよく表していると思われる。まず、単語レベルの文型パ

---

るにつれて、非線形構造が増加するとなると、あらかじめ準備しておくべき文型パターン数が禁止的な数に増大し、方式そのものが成り立たなくなるおそれがある。



ターンに比べて句レベルの文型パターンの「完全一致」が多いのは、以下の理由によると考えられる。すなわち、句レベルの句変数には元の標本文中の句を変数化しただけでなく、単語変数の適用範囲を句レベルに拡大したものなどがあり、入力文の中に元の標本文にはないような名詞修飾や格要素が入力文にあっても、それらは名詞句、動詞句などの一部として解釈されるためである。これに対して節レベルの文型パターンの「完全一致率」がそれほど高くないのは、文型パターン数そのものが少ないためと考えられる。

なお、節レベルの文型パターン数が少ないのは、節レベルで文型パターン化できない標本文が多かったためであるが、これは、多くの文が節レベルでは非線形であり、複数の節に分離して翻訳してその結果を合成するような要素合成法の方法では良い翻訳ができないことを意味している。また、(4)にあるように混合レベルの被覆率は句レベルと大差ないこと、節レベルの文型パターンでは、高品質の翻訳は期待しにくいことから、節レベルの文型パターンには余り期待しない方が良さそうである。

## (2) 適合文型パターン数

文型パターンでは、「離散記号」などによりさまざまな「任意要素」の存在が認められている。そのため、入力文に適合する文型パターンにおいて、1つの文型パターンに適合する仕方は、必ずしも一通りとは言えず、複数の適合の仕方（解釈の多義）が発生する場合がある。これは、「完全一致」する文型パターン、「部分一致」する文型パターンに共通する現象である。

以下、その例を示す。なお、例では文型パターンにマッチしなかった原文の要素を《 》で示している。この要素は、文型パターンとの照合において、「原文任意記号（離散記号）“/”」に一致したと判定されたものである。

例1) 入力文:「人を押しのけながらゲートのほうへ進んだ。」

<文型1> #1[N1は]/ VP2ながら / VP3.kako。

適合1: 人を押しのけながら ゲートのほうへ進んだ。

適合2: 人を押しのけながら《ゲートの》 ほうへ進んだ。

適合3: 人を押しのけながら《ゲートのほうへ》進んだ。

<英語文型> (N1 | I) VP3.past VP2.

(凡例) # [ ]:「要素補完記号」

例2) 入力文:「春の遠足は学年ごとに違う場所に行った。」

<文型1> #1[N1は]/ #2[N3の]/ #4[AJ5]/ N6に / V7.kako。

適合1: 春の 《遠足は学年ごとに違う》場所に 行った。

適合2: 春の 《遠足は》 学年ごとに《違う場所に》行った。

<英語文型> (N1 | I) V7.past to #2[N3.poss] #4[AJ5] N6.

<文型2> NP1は/ NP3に / VP4.kako。

適合1: 春の遠足は 学年ごとに違う場所に 行った。

適合2: 春の遠足は 学年ごとに 違う場所に行った。

適合3: 春の遠足は 学年ごとに 《違う》場所に行った。

<英語文型> NP1 VP4.pft in NP3.

【説明】適合文型2では他にも6通りの適合の仕方が存在するが、いずれも対応する英語文型パターンは翻訳に不適切である。

元来、同一の文型パターンに正しく適合する仕方は、高々1通りのはずであるから、上例のような場合は、複数の適合の仕方の中から、どれが正しい適合の仕方であるかを判定する方法が問題となる<sup>17</sup>。そこで、適合する文型パターンについては、文型パターン当たり、何通りの適合の仕方があるかについて評価した。

表8に、適合文型パターンを1つ以上持つ入力文を対象に、1文あたり適合した文型パターンの数と適合した文型パターン1つあたりの適合の仕方の数を示す。

表中の「延べ適合文型パターン数の平均」は、1入力文に対する適合文型パターン数の総和で、同一の文型パターンに複数の適合の仕方があった場合は、適合の仕方の数をカウントアップしている。「適合文型パターン当たりの解釈の数」は(前者)/(後者)で計算される。なお、混合レベルは3レベルの和である。この表から以下のことが分かる。

- (1) 単語レベルでは、入力文あたり適合する文型パターン数が少ない。また、文型パターン当たりの解釈の数は最も少ない。
- (2) 句レベルの異なり文型パターン数は、単語レベルの約5倍も多く、文型パターン当たりの解釈の数も2倍以上である。
- (3) 節レベルの場合は、異なり文型パターン数は少ないが、文型パターン当たりの解釈の数は最も多い。

単語レベルで適合文型パターン当たりの適合の仕方(解釈多義)が1.44件であることから、文法的レベルの文型パターンでも、単語レベルでは、かなり個別性の高い文型パターンとなっ

<sup>17</sup> 異なり文型パターンの中には、意味的に正しいものが複数存在することがある。また、それらに対応する英語文型が異なるときは、微妙な意味の違いによる訳し分けの対象となる。

表 8: 入力文 1 文に対する適合文型パターン数 (適合文型パターンを持つ入力文の場合)

文型パターン種別	延べ適合文型パターン数の平均	異なり文型パターン数の平均	適合文型パターン当たりの解釈の数
単語レベル	14.05 件	9.79 件	1.44 件
句レベル	164.72 件	51.48 件	3.20 件
節レベル	41.37 件	9.85 件	4.20 件
混合レベル	220.11 件	71.09 件	3.10 件

ていることが分かる。句レベル、節レベルになるに従って、文型パターンの解釈の多義が増加していることから、これらのレベルでは、今後、意味的な制約など、より強力な適合条件を指定することが必要と考えられる。

### 4.3 外国人用日本語例文に対する被覆率

前節まででは、文型パターン作成に使用した例文を対象に「自己パターン」以外の文型パターンにどれだけ適合するかを調べたが、本節では、参考のため文型パターン作成用の標本とは別の日本文を対象に被覆率を調べる。

試験に使用する例文は、「外国人のための日本語例文・問題シリーズ」全 18 巻 (名柄, 加藤, 福地 1989) に掲載された日本語例文から抽出したものである。具体的には、上記シリーズの例文 12,000 件に含まれる述部 2 又は 3 の重文と複文を取りだし、その中からそれぞれ 1/10 の割合で、801 件の例文をランダムに選択して試験用の入力文とした<sup>18</sup>。なお、以下では、この試験文を「外国人向け試験文」と呼ぶ。

#### (1) 被覆率特性

被覆率実験の結果を表 9 に示す。この結果を 4.1 節の (1) の結果 (表 5, 表 7: 「母集団試験文」の場合) と比較すると以下のことが分かる。

- (1) 「外国人向け試験文」では、単語レベル、句レベル、節レベルの「文型再現率  $R1$ 」と「文型一致率  $R2$ 」が共に 4~17% 低下しており、低下の割合は、単語レベルが最も大きい。

<sup>18</sup> 述部数 2 以上をすべて対象としたことでもあるが、このテキストは重文、複文の占める割合が大きい。これは、本テキストが外国人向けの教材であり、外国人の不得手な言い回しの例文が多いためと思われる。

- (2) 「完全一致」の割合は、3レベルとも10%~30%の範囲で低下している。相対値で見ると、単語レベルの低下は著しく、約1/3となっている。
- (3) しかし、「部分一致」の割合は、7~24%向上しており、「完全一致率」の低下が大きいものほど、「部分一致」の増加する割合が大きい。

表 9: 「外国人用試験文」に対する被覆率

文型パターン のレベル	文型再現率 (R1)			文型一致率 (R2)
	完全一致	部分一致	合計	
単語レベル	4.9 %	48.1 %	52.9 %	30.0 %
句レベル	24.7 %	58.2 %	82.9 %	64.5 %
節レベル	22.2 %	52.2 %	74.4 %	54.8 %
混合レベル	28.5 %	57.3 %	85.8 %	67.7 %

このうち、(1)で単語レベルの値が最も大きく低下するのは、このレベルの文型パターンは元となった標本文の母集団の個別的な特徴の多くを引きずっているのに対して、句レベル、節レベルでは、かなり一般化した文型パターンとなっていることを意味する。これは、単語レベルの完全一致率が大きく低下していること、また、文型パターン数が約1/10の節レベルの文型パターンでR1、R2共にあまり低下しないことから裏付けられる。

また、(2)と(3)で見られるように、「完全一致」する文型が減少することも、テスト文の母集団の違いが現れたものと思われる。「外国人向け日本文」では、外国人の苦手な日本語表現が多く含まれることが考えられるが、そのような表現こそ機械翻訳でも苦手な表現だと推定されるから、今後そのような表現を拾い出して文型パターン化することも大切だと考えられる。

(2) 入力文1文に適合する文型パターン数

適合する文型パターンがあった入力文に対して、1入力文にどれだけの文型パターンが適合したかを表10に示す。表の各欄の意味は、4.2節の(2)の場合(表8:「母集団試験文」の場合)と同様である。

この表と表8から以下のことが分かる。

- (1) 単語レベルでは、異なり文型パターン数、延べ適合文型パターン数共に、「母集団試験文」の場合に比べて、ほぼ半減している。
- (2) それに対して、句レベルの減少は30%程度にとどまっており、節レベルでは、殆ど低下していない。

表 10: 入力文 1 文に対する適合文型パターン数

文型パターン種別	異なり文型パターン数の平均	延べ適合文型パターン数の平均	適合文型パターン当たりの解釈の数
単語レベル	4.53 件	7.71 件	1.70 件
句レベル	33.73 件	120.10 件	3.56 件
節レベル	9.75 件	34.85 件	3.57 件
混合レベル	48.01 件	162.66 件	3.37 件

(3) 適合文型パターン当たりの解釈の数は、「母集団試験文」の場合と目立った違いはない。

これらは、単語レベルの文型パターンは汎用性が乏しく、母集団が異なる入力文に対して弱い。句レベル、節レベルになるにつれて汎用性が増大することを意味している。同一母集団のテスト文では、節レベルの文型パターンの有用性が乏しかったが、異なる母集団の入力文で効果を発揮することも予想される。

## 5 文型パターン記述方式と被覆率の関係

4.1 節の評価結果から見て、現状の文型パターンの被覆率は、実用上、まだ十分とは言えない。現在の文型パターンは、主として対訳例文の線形な要素を変数や関数に置き換えることによって作成されているが、線形な要素でも訳文生成で役立つ要素については、「任意要素」として文型パターンに記述されている。また、助詞、助動詞は表記のレベルで関数化されている。しかし、異なる時制や様相への適用は考えられていないこと、日本語に顕著な語順の自由度も考慮されていないことなど、種々の改良すべき余地を残している。

そこで、本章では、文型パターンの記述方法の違いによる被覆率特性の違いを明らかにするため、「任意要素指定が被覆率に与える効果」と「時制、相、様相の情報が被覆率に与える影響」についての評価実験を行った。

### 5.1 任意要素指定機能の効果

文型パターンの記述で使用される「任意要素」は、「原文任意要素」（「離散記号」で指定）と「文型任意要素」（「省略記号」で指定）に分けられる。

前者は、任意の文型要素が存在しても良い位置を表すものである。該当する位置に現れた入

力文の要素は、別途翻訳して英文に組み込まなければならないが、組み込むべき位置を知るためには構文解析処理などが必要である。これに対して、後者は、英文中に訳出すべき位置情報が指定されているため、該当する部分の翻訳を組み込むことは容易である。いずれも、単語レベル、句レベルの文型パターンの被覆率向上を目指したものである。

そこで、文型パターンの定義において、これらの要素の指定機能によってどれだけの汎化効果が得られたかを評価するため、任意要素指定機能の有無と被覆率の関係について「母集団試験文」1万件を対象に実験的に調査した。実験では、元の文型パターンに対して例に示すような変更を加えた辞書を使用した。

#### 例) 任意要素を削除した文型パターンの例

元の文型パターンから、1「原文任意要素(離散記号“/”)」と2「文型任意要素」を削除した文型パターンの例を示す。

<元の文型1> #1[REN2]N3をV4.rareru/のはN5しか/いない。

(参考例文): この事態をおさめられるのは彼しかいない。

1を削除 #1[REN2]N3をV4.rareruのはN5か/いない。

2を削除 /N3をV4.rareru/のはN5しか/いない。

1 2を削除 N3をV4.rareruのはN5しか/いない。

<元の文型2> #1[N1は]/[N2の]/N3をV4.teiru/(うちに|内に)/[N1は]/VP5.tekuru.kako。

(参考例文): 彼の言い訳を聞いているうちに怒りがこみ上げてきた。

1を削除 #1[N1は][N2の]N3をV4.teiru(うちに|内に)[N1は]/VP5.tekuru.kako。

2を削除 N3をV4.teiru/うちに//VP5.tekuru.kako。

1 2を削除 N3をV4.teiruうちにVP5.tekuru.kako。

(凡例)(うちに|内に):「要素選択記号」で、指定された複数要素のいずれかにマッチすることが要求される。

【説明】“#1[N1は]”は補完要素記号，“[N2の]”は任意要素記号で、いずれも2に該当する。また，“(うちに|内に)”も任意記号の1種であり、2の削除では、選択機能を停止し、第1要素へのマッチのみOKとする。

評価結果を表11に示す。表では、備考欄に上例との対応を示した。

この表から以下のことが分かる。

- (1) 「離散記号」(「原文任意要素」)の有無を比較すると、単語レベルでは、この記号の使用によって「文型再現R1」は5倍近く向上しており、句レベルでも1.5~2倍近く向上している。
- (2) 「文型任意記号」も被覆率の向上に与える効果は大きいですが、単語レベルでは、1.5倍程度、句レベルでは、1.1倍程度で「離散記号」の場合に比べてその効果は少ない。

これらから、「離散記号」は被覆率向上で大きな効果を持つことが分かる。しかし、適合した文型パターンを見ると、確かに、意味的に適切な文型パターンへの適合も増加しているが、不適切な文型パターンへの適合が予想を超えて大幅に増大しているようである。これは、「原文任

表 11: 任意要素指定の効果

場合分け		文型再現率 (R1)		備考 (例との対応)
離散記号	文型任意記号	単語レベル	句レベル	
あり	あり	69.8%	89.0%	元の文型
	なし	47.6%	83.7%	2 を削除
なし	あり	15.4%	55.6%	1 を削除
	なし	10.7%	48.7%	1 2 を削除

意記号」が任意の要素の存在を認める意味で使用されていることが原因と考えられる。現状の文型パターンでは、汎用性を向上させるため、入力文が文法的にも意味的にも正しいことを前提に大変緩やかな適用条件とされているが<sup>19</sup>、今後は、「原文任意要素」を文法的、意味的に分類し、制約条件を付与すると共に、その挿入位置を見直すことなどにより、間違った文型パターンへの「部分一致」を防止することが大切と考えられる。

これに対して、「文型任意記号」の効果は、今ひとつ少ないように思われる。これは、半自動的な方法で「文型任意要素」の指定が行われているため、字面表記の揺らぎが十分に吸収されていないことが予想される。今後は、徹底した異表記のグループ化を行うと共に、部分的な言い換え表現の辞書を開発することなどが期待される。

## 5.2 時制・様相などの縮退効果

単語レベルの文型パターン化において、時制、相、様相の表現は原則として関数化されているが、「文型任意要素」としての指定は行われていない。文型パターンの照合段階での時制変形、相変形、様相変形などの操作を認め、そのような変形が可能な関数を「文型任意要素」として指定すれば、文型パターンの汎用性は大幅に向上することが期待される。そこで、現在使用されている時制、相、様相の関数を無視した場合、文型被覆率がどのように変化するかを調査した<sup>20</sup>。

この実験は、文型パターン照合プログラムの照合条件を以下のように変更することによって行った。すなわち、入力文内の述語末尾の助詞、助動詞が、文型パターン内の時制、相、様相を表す関数の条件を満たさなくても文型パターンは適合したものと判定した。これにより、例えば、下記の文型パターンとの照合では、AJ5の末尾に *.hitei^shushi.souda* に適合する表現が

<sup>19</sup> 現在「離散記号」は、その位置に任意の要素が挿入されても良いことを意味する記号として、文型パターン内のほぼすべての文節境界に付与されている。

<sup>20</sup> 効果を推定使用とする方法は、例えば、過去形の例文や推量の意味の例文から、原型に縮退させた文型パターンを作成し、翻訳時には、この文型から過去形、未来の表現、推量や否定などの表現を生成する方法である。この方法は、非線形な要素に対しては適用できない。本節の検討は、すべての削除できたときの効果を調べるもので、効果の最大値を知ることが目的としている。

なくとも、入力文は文型パターンに適合したものと判定される。

例)  $N1$  が/不便なため、/ $N2$  の/ $N3$  は/#4[あまり]/AJ5[.hitei^shushi.souda]。

(適合する入力文の例) 交通が不便なため、土地の値段はあまり高くないそうだ。

【説明】入力文は、文型パターン末尾の関数部分 “.hitei^shushi.souda” の適合条件を満たさなくても良いとすることにより、入力文の文末が「高くない」、「高い」、「高かった」、「高くなかった」、「高いようだ」、等多くの場合に適合するようになる。

実験では、文型パターンの作成で使用した 12.9 万件の日本語（「母集団試験文」）を入力文として、単語レベルの文型パターン辞書 12.2 万件と照合し、文型再現率を求めた。その結果を表 12 に示す。また、文型一致率と適合文型パターン数の平均を表 13 に示す。

表 12: 時制・相・様相の縮退効果 (単語レベル)

条件	完全一致	部分一致	文型再現率 (R1)	文型一致率 (R2)
削除前	14.4%	50.3%	64.7%	42.08%
削除後	6.6%	91.8%	98.4%	64.49%

表 13: 文型一致率 (R2) と一致文型数平均

条件	一致文型パターン数平均 / 入力文		適合文型パターン
	異なり文型パターン数	延べ文型パターン数	当たりの解釈の数
削除前	9.45 件	22.92 件	2.43
削除後	73.5 件	255.8 件	3.48

表 12, 表 13 の「削除前」は現状の文型パターンの場合を示すが、その値は、表 7 の値と若干異なる値となっている。これは、テストに使用した入力文が異なるためである（いずれも「母集団試験文」であるが、テスト文は、表 7 の場合 1 万文、表 12, 表 13 の場合 12.9 万文）。

表 12 から、時制、相、様相の削除によって、「完全一致率」は大きく減少する<sup>21</sup>が、「部分一致」が大幅に増えて、100%近い「文型再現率 R1」になり、被覆率向上で大きな効果が期待できることが分かる。また、表 13 からは、削除後の「文型一致率 R2」は 1.5 倍程度に向上するが、その反面、1 入力文に対する適合文型パターン数とその解釈の数は 10 倍程度に増加することが分かる。

<sup>21</sup> これは、現在形の基本形式のテスト文しか文型と完全一致しないようになるため、当然の結果である。



これらのことから、今後、任意化により、時制、相、様相の情報を極力縮退させることができれば大きな効果が期待される。しかし、その場合は、不適切な適合文型パターンも大幅に増大することが確実である。時制、相、様相の情報は、対応する英語文型パターンを決定する上で重要な情報であり、非線形要素となる場合も多いと予想される。任意化の可否判定規則を検討することにより、不適切な文型パターンへの適合が抑止できるように改良することが大切だと考えられる。

## 6 排他性の評価

前章まででは、文法レベルで記述された文型パターンについて「文型再現率  $R1$ 」と「文型一致率  $R2$ 」を用いて「再現性」を評価した。しかし、適合した文型パターンが果たして意味的に適切な文型パターンであるか、また、それに対応する英語文型パターンが訳文の生成に問題なく使用できるかは大変重要である。そこで、本章では、排他性が最も高いと推定される単語レベルの文型パターン対象に排他性に関する評価を行った<sup>22</sup>。

### 6.1 評価の方法

まず、文型パターン作成に使用した例文 12.9 万件（「母集団試験文」）に対する文型パターン照合実験の結果から、「自己パターン」以外に 1 件以上の適合文型パターンを持つ入力文を抽出した。次にその中から、ランダムに 1/400 のテスト文（177 文）を選び、それらに適合する文型パターン（「自己パターン」を除く 1,073 件）について表 14 に基づき正誤判定を行った。

表 14: 評価の基準

適合文型 パターンの品質	評価の区分	用途
ランク 1	最低限入力文の 構文構造と一致する	英語文型パターンが必ずしも翻訳に使用できるとは限らない。翻訳に使用できない場合は、従来の方法で翻訳しなければならないが、文型パターンに構文情報（係り受け構造の情報など）を付与しておけば、改めて構文解析をしなくてすむ。
ランク 2	意味的にも正しく 英語文型が翻訳に使える	対応する英語文型パターンを使用して翻訳ができる。

備考 ランク 1 の文型パターンは、ランク 1 の文型パターンも含む点に注意

<sup>22</sup> この評価は、適合した多数の文型の意味を判定しなければならず、多大なコストがかかるため、もっとも排他性の高いことが期待される単語レベルに限ることとし、標本数も限定して評価した。

評価基準として2つのランクを設けたのは以下の理由である。すなわち、文型パターン辞書の本来の目的は、入力文の翻訳に使用できる目的言語の文型パターンを求めることであり、そのような文型パターンがないときは従来の翻訳方式で翻訳することになるが、原言語側の文型パターンにあらかじめ構文情報を持たせていけば、従来方式で翻訳するに際して、構文情報が役に立つと期待されるためである。

入力文と適合した文型パターンの日本語側を見ると、係り受け構造や格要素の意味役割（主格，目的格など）が同じで、意味的にも正しいと思われる場合でも、英語文型パターンが変型の内容に依存する場合や、英語文型パターンの骨格が英語の訳語に依存している場合は、英語文型パターンが使えないため、ランク1の分類とした。入力文に対してランク1で適合する文型パターンとランク2で適合する文型パターンの例を示す。

例1)「老婆は髪を振り乱しながら賊を追いかけた。」

<ランク1で適合する文型パターン>

#6[N1は]/N2を/V3ながら/N4を/V5<sup>renyou.kako</sup>.

(N1|I) V5.past on (N1|I).pron.poss N4 while V3.ing the pictures in N2.

【説明】 適合文型パターンは構文構造も入力文と一致するが、対応する英語文型パターンには、日本語文型パターン内で対応付けがされていない要素として“the pictures in”という字面が使われている。そのため、この入力文の翻訳には使用できない。この文型パターンは、下記のような例文ではランク1で適合する。

「アルバムを見ながら学生時代を回顧した。」

I thought back on my student days while looking at the pictures in the album.

<ランク2で適合する文型パターン>

N1は/N2を/V3ながら/N4を/V5.kako. N1 V5.past after N4, N1.pron.poss N2 V3.ing.

(参考) 英語文型パターンを使用した翻訳結果

The old woman chased after the thief, her hair flying in disarray.

例2)「彼はマフィアの大親分だと言われている。」

<ランク1で適合する文型パターン>

N1は/#2[AJ3]/N4.daと/V5.rareru.teiru. N1 be V5.ed as #2[AJ3]N4.

【説明】 “~ be regarded as” のように「be 過去分詞 as」と訳出して良い動詞は限られており、この英語パターンを用いて英語に訳出しようとするとき“He is said as ~”となり誤った英語になってしまう。この文型パターンは、下記のような例文ではランク2で適合する。

「彼はよい教師だと思われている。」 He is regarded as a good teacher.

$N1$  は/ $N2$  の/ $N3$ .*da* と/ $V4$ .*rareru.teiru*.  $N1$  be  $V4$ .*ed* to be  $N2$ 's  $N3$

【説明】 上例と同様， $V5$  を “~be considered to be” で訳せる場合は限られている．この文型パターン下記の入力文にはランク 2 で適合する．

「ポロは金持ちの娯楽だと考えられている。」

Polo is considered to be a rich person's pastime.

<ランク 2 で適合する文型パターン>

$N1$  は/ $N2$  の/ $N3$ .*da* と/ $V4$ .*reru.teiru*.  $N1$   $V4$  to be  $N3$  of  $N2$  family.

(参考) 英語文型パターンを使用した翻訳結果

He is said to be the godfather of a Mafia family.

## 6.2 評価結果

適合文型意味正解率 ( $P1$ )

「適合文型意味正解率」の評価結果を表 15 に示す．表中「完全一致」「部分一致」の欄は，入力文に「完全一致」又は「部分一致」した文型パターンの中で，それぞれのランクに適合した文型パターンの割合を表す．また「全体」の欄は，入力文に「完全一致」もしくは「部分一致」した文型パターン全体が対象で，それぞれのランクに相当する文型パターンの数の割合，すなわち「適合文型意味正解率  $P1$ 」を示す．

表 15: 適合文型意味正解率  $P1$

評価のレベル	完全一致	部分一致	全体 (適合文型意味正解率 $P1$ )
ランク 1	77.3%	11.8%	18.4%
ランク 2	41.1%	5.4%	9.1%

この表から「完全一致」した文型パターンに比べて「部分一致」した文型パターンの品質は極端に悪いことが分かる．4章の結果では，単語レベルの文型パターンの場合「完全一致」する文型パターンに比べて「部分一致」する文型パターンは大変多い．多数の文型パターンの中からきわめて少数の意味的に正しい文型パターンを選択することは難しいと予想される．しかし「意味的正解率」が小さいとは言え，多数の「部分一致」の文型パターンの中にはそれなりに正しい文型パターンが含まれることになるから「部分一致」した文型パターンを無視して良いかどうか微妙である．いずれにしても不正解な「部分一致」文型パターンが多数適合することが問題である．この問題の最大の原因は，5.1節で述べたように「原文任意要素」を指定する「離散記号」の使用基準にあると思われる．

適合文型正解含有率 ( $P_2$ )

文型パターンを使用した翻訳では、通常、各入力文に対して使用できる文型パターンが1つ以上存在することが望まれる。そこで、適合文型パターンがあった入力文を対象に「適合文型正解含有率  $P_2$ 」を評価した結果を表 16 に示す。表中の「完全一致」の欄は、「完全一致」した文型パターン（いずれも1つ以上）を持つ入力文が対象で、1入力文に「完全一致」した文型パターンの中に各ランクに該当する文型パターンが1つ以上含まれる割合を示す。これに対して、「部分一致」の欄は、「部分一致」した文型パターンしか持たない入力文が対象で、適合文型パターンの中に該当する文型パターンが1つ以上含まれる割合を示す。また、「全体」の欄は、適合した文型パターン（「完全一致」、「部分一致」のいずれでも良い）の中に、該当する文型パターンが1つ以上含まれる割合、すなわち「適合文型正解含有率  $P_2$ 」を示す。

表 16: 適合文型正解含有率  $P_2$

評価のレベル	完全一致	部分一致	全体 (適合文型正解含有率 $P_2$ )
ランク 1	81.0%	17.8%	35.6%
ランク 2	50.0%	9.6%	20.9%

この表から、適合した文型パターン全体の「適合文型正解含有率  $P_2$ 」は、「適合文型意味正解率  $P_1$ 」(表 15) に比べて、約 2 倍大きいことが分かる。

正解率に関するまとめ

以上、単語レベルの適合文型パターンの意味的な正しさに関する評価結果から以下のことが分かる。すなわち、適合文型パターンの中から翻訳に使用する文型パターンを選択するとき、

- (1) ランダムに文型パターンを選べば、それが意味的にも適切な確率は 9% である。
- (2) もし最適な文型パターンを選ぶ方法が分かった時は、意味的な正解率は 21% となる。

なお、

- (3) 適合文型パターンのうち、「完全一致」した文型パターンの適合率は、「部分一致」した文型パターンの場合の 5~10 倍大きく、「完全一致」した文型パターンからランダムに選べば 41%、意味的に正しい文型パターンを選択することができれば 50% である。

以上から、単語レベル文型パターンの場合、適合した文型パターンから意味的に正しい文型パターンと誤った文型パターンを区別する方法が実現できれば、入力文の約 14% ( $=0.698\% \times 21\%$ ) に対してに適切な文型パターンが発見できること、またその場合、各入力文の  $2/3$  の部分に対応する英語文型パターンをそのまま使えば翻訳できることが分かる。

## 7 評価結果のまとめと考察

前章までの評価結果をまとめ、今後、文型パターンの適合率と正解率を向上させる方法について考察する。また、文型パターンを使用した機械翻訳の可能性について考える。

### 7.1 評価結果のまとめ

#### (1) 文型パターン数と「再現率」の関係

重文、複文の文型パターン化において、標本文が1万件以下の場合には、良い「再現率」は得られないが、逆に、標本文数を数万件以上に増大させても飽和現象が発生し、その割に「再現率」は向上しない。このことから、現状の文型パターン数12万件(単語レベルの場合)を基本に、今後その改良を図ることが現実的と考えられる。

#### (2) 文型パターンの「再現率」と「意味適合率」の問題

単語レベル、句レベル、節レベルの文型パターンの「文型再現率  $R1$ 」は、それぞれ、70%、89%、78%であり、「文型一致率  $R2$ 」は、47%、76%、63%である。但し、「外国人向け試験文」では、10%あまり減少する。

このように多くの入力文に対して適合する文型パターンが存在するが、意味的に見た場合、入力文の翻訳に適用できる文型パターンの割合(21%)は小さく、現状では、実用に耐えられる品質とは考えられない<sup>23</sup>。これが文法レベルで記述された文型パターンの限度と思われることから、今後は、文型パターンに意味レベル的な制約条件を追加することが重要と考えられる。

#### (3) 節レベルの文型パターンの再現率

節レベルの汎化ができる標本文は少ないため、全体から見ると、その「再現率」は、単語レベル、句レベルの文型パターンに比べて比較的小さい。従って、節レベルの文型パターンの役割は保険として捉え、再現率向上の点では、句レベルの文型パターンに重点を置いて今後の汎化を進めるのが適切と判断される。

#### (4) 適合文型パターンを使用した機械翻訳

単語レベルの文型パターンでは、「部分一致」する文型パターンに比べて「完全一致」する文型パターンの割合は低い。また句レベルの文型パターンでも適合する文型パターンのかなりのものが「部分一致」である。意味的に不適切な文型パターンに「部分一致」する原因としては、「離散記号」の使用基準が緩すぎる事が挙げられる。今後は、「離散記号」に制約条件を加え、文型パターン内での指定位置を見直すことが大切と考えられる。

<sup>23</sup> 単語レベルだけでなく句レベル、節レベルの文型パターンの意味的な被覆率にもよるが、単語レベルの文型パターンでも現状(14%)の2倍位は欲しいところである。

### (5) 意味的に適切な文型パターンの選択

適合文型パターンを持つ入力文について見ると、適合する文型パターン数(3~60件)は大変多い。また、適合する文型パターン一件当たりの適合の仕方(単語レベルで平均14件、句レベルで165件)も多い。多数の適合文型パターンの中から意味的に適切な文型パターンとその解釈を決定するのは大変難しい問題であると予想される。今後、極力、意味的に不適切な文型パターンには適合しないような工夫が重要である。

## 7.2 今後の対策について

既に述べたように文型パターン方式の成否の鍵は、どれだけの被覆率と意味的排他性が実現できるかであるが、被覆率について明確な成否の分岐点を設けるのは難しい。文型パターンは、非線形な構文構造を対象に作成されているが、線形で従来の翻訳方法でも十分訳せるような入力文にも適合する。本研究では、単語レベル、句レベル、節レベルを合わせた被覆率(この場合「文型再現率  $R1$ 」×「適合文型正解含有率  $P2$ 」<sup>24</sup>で評価)として50%が得られれば、実用上の効果が期待できると考え、それ以上の被覆率の達成を当面の実現目標としているが、現状(文法レベル文型パターン)の「適合文型正解含有率  $P2$ 」の値は低く、適合した文型パターンから意味的に正しいものを選択することも容易でないと推定される。また、意味的排他性も実現されていない。従って、今後の課題は、第1に「文型再現率  $R1$ 」と「適合文型正解含有率  $P2$ 」の両者を上げること、第2に「文型パターンの意味的排他性を向上させ意味的に不適切な文型パターンに適合しないようにすること」の2点であると考えられる。

そこで、以下では、これらの2つの課題の可能性について考える。

### 意味的な排他性向上のための課題について

上述の通り、適合文型パターンの中から意味的に正しい文型パターンを選択することはかなり困難な仕事だと思われる。この問題を解決する方法としては、文型パターンで使用された変数に意味的制約条件を付与することが有効と見られる。そこで、どの程度の分解能で意味属性を付与すると良いかであるが、本検討の結果から見て、文型パターンでは、平均数十通りの意味が縮退していると思われるから、平均で数十通り、重なり大きいところを考えても1,000通りくらいの分解能(「日本語語彙大系」(池原他 1997)の意味属性の分解能のレベルに相当)があれば、ほぼ意味的な多義は解消できそうである。

<sup>24</sup> 適合文型パターンの品質はランク2での評価とする。

## 被覆率向上のための課題について

変数に意味的制約条件を付与した場合、「意味適合率」は大幅に向上すると期待できるが、逆に「文型再現率  $R1$ 」と「文型一致率  $R2$ 」は、大幅に低下することが予想される。従って、これらの値を大幅に向上させる方法が必要である。

その対策であるが、「日本語語彙大系」の開発例と今回の被覆率調査から見て(1)「格要素の語順の自由指定や副詞の位置変更可能指定を行うこと」、(2)「時制、相、様相の汎化を行うこと」、また、(3)「字面表記の自立語、付属語の表記の揺らぎを徹底的にグループ化すること」が重要と考えられる<sup>25</sup>。

そこでまず(1)について考えると、格要素の語順変更指定の効果は、単語レベルの文型パターンから句レベルの文型パターンに汎化した場合と同程度の被覆率向上が見込めそうで、適合文型パターンは何倍かに増加する可能性がある。

次に、(2)の効果であるが、現在使用されている時制、相、様相の情報を汎化した場合についての効果推定のための実験結果を見ると、単語レベルの「文型一致率  $R2$ 」が、54%から最大98%に向上する可能性が示されている。実際に汎化できるのは線形要素に限られるが、それでも適合文型パターンはかなり大幅に増大することが予想される。

第3に(3)の効果であるが、実験的評価において、現在の文型パターンでは、「離散記号」を使用することにより、「文型一致率  $R2$ 」は、単語レベルで4~5倍、句レベルで1.5倍~2倍に向上したことが示されている。また、「文型任意記号」も単語レベルでは、数十%の効果を持つことが示されている。これらから、より徹底した汎化を行えば、まだ相当の効果が得られるものと期待できる。

## 8 あとがき

日本語重文、複文に対する文法レベルの文型パターン辞書(単語レベル12.2万件、句レベル8.8万件、節レベル1.1万件)を対象に、被覆率特性(「文型再現率」、「文型一致率」、「適合文型意味正解率」、「適合文型正解合有率」の4種)を調べ、今後の課題と文型パターン方式の可能性について検討した。

その結果によれば、現状の文型パターンの「文型再現率  $R1$ 」は、単語レベル、句レベル、節レベルの順に70%、89%、78%で、いずれもかなり高いこと、また、単語レベルの文型パターンが「完全一致」する割合(15%)は、句レベル(54%)、節レベル(40%)に比べてかなり小さいことが分かった。「部分一致」した文型パターンを使用した翻訳処理より「完全一致」した文型パターンを使用した翻訳処理の方が容易であること、また、後者の文型パターンの方が意味的に正しい割合が大きいことから、機械翻訳では、単語レベルで完全一致する文型パターン

<sup>25</sup> 「日本語語彙大系」の結合価パターンの開発では、これらの作業に加えて変数意味属性自身の汎化作業も徹底して行っており、大きな効果のあることが確認されている。しかし、精密な意味属性汎化は機械化が困難である。今回の文型パターン化では、文型パターン数も多く、膨大な人手コストがかかると予想されるため、意味属性の汎化においても、なるべく機械的な方法の実現が望まれる。

がない場合は、句レベルや節レベルの文型パターンを使用するなど、各レベルの文型パターンを組み合わせる使用することが重要と考えられる。

また、適合文型パターンに対する意味的な正解率の評価では、単語レベル文型パターンの場合、適合文型パターンの中に意味的に正しいものが含まれている確率が、20%弱にとどまっております。入力文は多数の意味的に適合しない文型パターンと適合してしまうことが分かった。これは、文法レベルの情報で記述された文型パターンの限界を示しているものと見られる。

そこで、今後は、1 意味的に適切な入力文には、なるべく漏らさず適合するよう文型パターンを汎化すること、また、その際、2 可能な限り意味的に正しいもののみが掬い取れるような仕組みを考えることの2点を目標に、意味レベル文型パターン辞書を開発する計画であるが、このうち、1の課題としては、格要素の語順の自由指定や副詞の位置変更可能指定を行うこと、時制、相、様相の汎化を進めること、また、字面表記の自立語、付属語の表記の揺らぎを徹底的にグループ化すること等が挙げられる。今回の実験結果から、特に時制、相、様相の汎化の効果が大きいと予想され、これらによって意味的に適切な文型パターンの割合は数倍程度以上に向上することが期待される。また、2では、文型パターン内の変数の変域に対して意味的制約条件を付けることが必要と考えられるが、現在の多義の発生程度から考えると、その意味的分解能としては1,000種類の分類があれば良さそうである。

以上、本研究では、文型パターン辞書開発上の様々な問題についての指針が得られたので、今後は、これらの指針に基づき、意味レベルの文型パターン辞書の開発に取り組んでいきたい。

#### 謝辞

この研究は、科学技術振興機構（JST）の戦略的創造研究推進事業（CREST）で行っているものである。技術会議等を通じてご議論を頂いた宮崎正弘先生（新潟大）、佐良木昌先生（長崎純心大）、池田尚志先生（岐阜大）、新田義彦先生（日本大）、山本和英先生（長岡技科大）、横尾昭男様（NTT-AT社）に感謝する。また、文型パターン化作業を担当して頂いたNTTアドバンステクノロジー株式会社の皆様に感謝する。

#### 参考文献

- 有田潤 (1987). ドイツ語講座 II. 南江堂.
- 市川亀久彌 (1963). 創造的研究の方法論 (増補版). 三和書房.
- 池原悟, 宮崎正弘, 白井諭, 林良彦 (1987). “言語における話者の認識と多段翻訳方式.” 情報処理学会論文誌, 28 (12), 1269–1279.
- 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (1997). 日本語語彙大系. 岩波書店.
- Ikehara, S. (2001). “Meaning Comprehension Using Semantic Patterns in a Large Scale Knowledge-Base.” In *Proceedings of the PACLING'01*, pp. 26–35.



- 池原悟 (2001). “自然言語処理の基本問題への挑戦.” 人工知能学会誌, 16 (3), 422–430.
- 池原悟 (2002). “究極の翻訳方式の実現に向けて = 類推思考の原理に基づく翻訳方式 = .” *AAMT Journal*, アジア太平洋機械翻訳協会, 32, 1–7.
- 池原悟, 宮崎正弘, 佐良木昌, 池田尚志, 白井諭, 村上仁一, 徳久雅人 (2003). “機械翻訳のための日英文型パターン記述言語.” 電子情報通信学会技術研究報告, 「思考と言語」研究会, TL2002-48, 1–6.
- 池原悟 (2003). “言語で表現される概念と翻訳の原理.” 電子情報通信学会技術研究報告, 「思考と言語」研究会, TL2003-25, 7–12.
- 池原悟 (2004). “非線形な言語表現と文型パターンによる意味の記述.” 情報処理学会研究報告, 自然言語処理研究会, 2004-NL-159, 139–146.
- 池原悟, 阿部さつき, 徳久雅人, 村上仁一 (2004). “非線形な表現構造に着目した日英文型パターン化.” 情報処理学会研究報告, 自然言語処理研究会, 2004-NL-160, 49–56.
- 金出地真人, 池原悟, 村上仁一 (2001). “結合価文法による動詞の訳語選択能力の評価.” 情報処理学会第 63 回全国大会, 2 巻, pp. 267–268.
- 名柄迪, 加藤康彦, 福地努 (編) (1989). 外国人のための日本語例文・問題シリーズ. 荒竹出版. 全 18 巻.

## 略歴

- 池原 悟: 1967 年大阪大学基礎工学部電気工学科卒業. 1969 年同大学大学院修士課程修了. 同年日本電信電話公社に入社. 数式処理, トラフィック理論, 自然言語処理の研究に従事. 1996 年スタンフォード大学客員教授. 現在, 鳥取大学工学部教授. 工学博士. 1982 年情報処理学会論文賞, 1993 年同研究賞, 1995 年日本科学技術情報センター賞 (学術賞), 同年人工知能学会論文賞, 2002 年電気通信普及財団賞 (テレコム・システム技術賞) 受賞, 電子情報通信学会, 人工知能学会, 言語処理学会, 機械翻訳協会各会員.
- 徳久 雅人: 1995 年九州工業大学大学院情報工学研究科博士前期課程修了. 同年同大学情報工学部助手. 統合的知能エージェントの開発に従事. 現在, 鳥取大学工学部助手. 自然言語処理の研究に従事. 情報処理学会, 電子情報通信学会, 人工知能学会, 言語処理学会各会員.
- 竹内 (村本) 奈央: 1999 年津田塾大学学芸学部英文学科卒業. 同年 NTT アドバンステクノロジー (株) に入社. 言語処理用辞書, ルールの開発に従事. 2004 年同社退社. 現在, フリーで言語処理用辞書を開発を行う. ACL, 日本英語学会各会員.
- 村上 仁一: 1984 年筑波大学第 3 学群基礎工学類卒業. 1986 年筑波大学修士課程理工学研究科理工学専攻修了. 同年 NTT に入社. NTT 情報通信処理研究所に勤務. 1991 年国際通信基礎研究所 (ATR) 自動翻訳電話研究所に出向. 現在, 鳥取大学工学部助教授. 主に音声認識のための言語処理の研究に従事.

電子情報通信学会，日本音響学会，言語処理学会各会員．

(1995年5月6日受付)

(1995年7月8日再受付)

(1995年9月10日採録)