非線形な重文複文の表現に対する文型パターン辞書の開発

池原 悟[†1]　徳久 雅人[†2]　村上 仁一[†1]
佐良木 昌[†2]　池田 尚志[†3]　宮崎 正弘[†4]

†1鳥取大学工学部　〒680-8552鳥取市湖山町南4-101　{ikehara,tokuhisa,murakami} ike.tottori-u.ac.jp
†2日本大学　〒101-0061東京都千代田区三崎町1-3-2　saraki@st.rim.or.jp
†3岐阜大学工学部　〒501-1193岐阜市柳戸1-1　ikeda@info.gifu-u.ac.jp
†4新潟大学工学部　〒950-2181新潟市五十嵐2の町8050　miyazaki@ie.niigata-u.ac.jp

あらまし　品質の良い機械翻訳や言い換え技術などを実現するため, 最近, 類推の原理による意味的等価変換方式が提案されている. この方式で必要とされる「意味類型パターン辞書」を実現するため, 重文と複文を対象に「文型パターン辞書」を開発した. 文型パターンは意味類型パターンと同様, 対象に対する人間の認識を表すための言語表現の枠組みであり, 同時に言語表現の意味をすくい取るための網の目のようなものである. 日英対訳文100万件から重文と複文の例文15万件を抽出し, 線形な要素を単語, 句, 節の順に汎化して単語レベル(12.3万件), 句レベル(8.0万件), 節レベル(1.2万件)の文型パターン(合計21.5万文型)を作成した. 汎化された自立語, 句, 節は, それぞれ62%, 22%, 4%であった. 汎化可能な句や節が少なかったことから, 要素合成法に基づく翻訳方式では, 対訳用例のような品質の良い翻訳はできないことが明らかとなった. これに対して, 全体の意味的な被覆率は71～77%であり, 文型パターン辞書を併用した翻訳方式の発展が期待される.
キーワード　文型パターン, 機械翻訳, 言い換え, 非線形表現, 意味類型, 類推翻訳, 重文複文

Development of Pattern Dictionary for Non-linear Structures of Complex and Compound Sentences

Satoru Ikehara[†1]　Masato Tokuhisas[†1]　Jin'ichi Murakami[†1]
Masashi Saraki[†2]　Takasi Ikeda[†3]　Masashiro Miyazaki[†4]

†1Tottori University,Tottori-city,680-8552 Japan　{ikehara,tokuhisa,murakami}@ike.tottori-u.ac.jp
†2 Nihon University,Tokyo,101-0061 Japan, saraki@st.rim.or.jp
†3 Gifu University,Gifu-city,501-1193 Japan, ikeda@info.gifu-u.ac.jp
†4 Niigata University,Niigata-city,950-2181 Japan, miyazaki@ie.niigata-u.ac.jp

Abstract　*Semantic typology based Analogical Mapping Method is expected as a method* of improving the quality of MT systems and rewording system. In order to realize a Semantic Pattern dictionary *(SMP-dictionary)* needed for this method, *Sentence Pattern dictionary (STP-dictionary)* for Japanese complex and compound sentences was developed. As is the case of *SMP, STP* is a framework to present human's conceptual cognition and therefore, is what looked like a net's mesh to scoop up the meaning of an expression. 150,000 of compound and complex sentences were extracted from the parallel copus of one million pairs for Japanese and English sentences and linear components were generalized in the order of full words, phrases and clause. The generated *STP-dictionary* (215,222 patterns) consists of *word-level senntennce patterns* (122,642 patterns), *phrase-level sentence patterns* (80,130 patterns) and *clause-level sentence patterns* (12,450 patterns). The ratios of generalized full word, phrases and clauses were 62%, 22% and 4.3%. The ratio of linear clause is astoundingly low. This shows that the high quality translations just like example sentences cannot be obtrained by MT methods based on *compositional semantics*. On the other hand, semantic coverage of *STP-dictionary* was 71-77 %. MT method with this dictionary will promise a great future.
Key word　Sentence Pattern, MT, Reeording, Non-linear Expression, Semantic Typology, Analogy

1. はじめに

1980年代以降, 機械翻訳の分野では大規模な投資が行われ大きく発展した[1]. しかし, 日英言語のような離れた言語間の翻訳は依然として困難で, その後も新しい方式に関するさまざまな研究が行われている. 中でも長尾によって提案され[2], 佐藤によって具体化された[3,4]用例翻訳は, 直接対訳例をまねる方法であるため解析の誤りが避けられるものとして着目された. しかし, きわめて大量の用例を必要とすること, また, 用例が増加するに伴って, 相互干渉の問題が避けられなくなることなどのため, 実用レベルに達していない.

これに対して, 最近では統計翻訳の研究が盛んである. この方式は, 同一言語族の英仏言語間の翻訳を対象に提案された[5]が, 音声認識の研究で成功した

HMMの技術が導入されてからは，日英翻訳など異なる言語族間の翻訳にも適用されている[6]．しかし，統計的な信頼性が保障できるだけの標本データを収集することは大変困難で，旅行会話など限定された適用範囲にとどまっている．

このように，用例データを直接使用する方法には限界があり，体系的に整理された言語知識ベースを実現することが重要と考えられる．

ところで，人間の知識を加え加工した知識ベースを使用する方法としては，古くからパターン翻訳がある[7-10]．この方式は，適合したパターンが見つかると高品質の訳文が生成できるため，トランスファー方式[2]や翻訳メモリと併用され[11,12]，すでに多くの商用システムに採用されている．しかし，用意されている文型パターン数はいずれもきわめて少数（高々200件）で，専門分野の特定の表現を対象としているに過ぎない．その理由としては，文型パターンの開発コストが高いことが挙げられるが，それ以上に，意味的な整合性の実現が難しいことが原因だと思われる．自動学習の研究では，文型パターンを自動生成する方法が研究されている[13-15]が，得られた文型パターンの品質に問題があること，また，出現頻度の小さい表現からは文型パターンが得られないことなどが問題で，実用にはまだまだほど遠い．

このような現状を克服することを狙って，かつて多段翻訳方式[16]が研究開発された．この方法は，話者の感情や意思を表現する主体的表現と対象のあり方を捉えた客体的表現を分離し，客体的な表現を対象に文型パターン辞書を作成することとすれば，必要な文型パターン数は大幅に減少するだろうと考えたものである．すでに，動詞と名詞の意味的な関係を対象に結合価パターン辞書（「日本語語彙大系」[17,18]）が実現され，その結果，単文の翻訳品質は飛躍的に向上したことが報告されている．

しかし，この方法では，まだ2つの問題が残されている[19,20]．一つは，入力文に対して単一の翻訳結果しか得られないことである．言語では，同一の内容を表すにもさまざまな表現の仕方がある．翻訳者は，複数の文型候補から文脈などを判断して最も適切な文型を選んで訳文を作成する．機械翻訳でもこのような訳し分けの機能を実現することが期待される．また，第2の問題は適用範囲の問題である．結合価パターンでは，重文や複文で表される事象間の関係構造の意味は表現されない．

これら2つの問題を解決する方法として，最近，類推原理による「意味的等価変換方式」が提案された[21]．この方式を実現するには，「意味類型パターン辞書」が必要である．この辞書は，非線形な言語表現構造に対する文型パターンを意味的に類型化したものである．

そこで，本研究では，「意味類型パターン辞書」の開発に先立って「文型パターン辞書」を開発したので報告する．以下，本論文では，「意味的等価変換方式」の概要を紹介したのち，「文型パターン辞書」開発の基本的な考え方と開発の結果について述べる．

2．意味的等価変換方式の概要

「意味的等価変換方式」は，与えられた言語表現に対して，それと意味的に等価な表現を発見して対応づける方法で，「意味類型論」と「等価的類推の原理」の二つの理論を背景としている．「意味類型論」は，「人間の対象認識は，認識論的な枠組みを用いて形成される」こと，また，「その枠組みとして母国語の表現の形式が使用される」ことを指摘したもので，ドイツ言語学者の有田潤によって提唱された[22]．また，「等価的類推の原理」は，「人間の独創的思考は，対象間における何らかの類推を背景としている」ことを指摘したもので，市川亀久弥が，「独創的研究の方法論」[23]の中で明らかにした．

「意味的等価変換方式」は，「意味類型パターン辞書」と「理論的意味範疇」から構成される．「意味類型パターン辞書」は，「意味類型論」に基づき，実際に使用された言語表現の中から，認識論的な思考の枠組みといえるような形式を取りだし，意味的にグループ化したものである．これに対して，「論理的意味範疇」は，類推の背景に存在する共通見地として市川が提案した概念を，意味類型パターンの間の意味的な対応関係を発見するための仕組みとして具体化したものである．中身は，言語表現の意味を定義するための概念（真理項と呼ぶ）を階層化したものである．意味類型パターンの意味は「真理項」を用いて定義される．

この方式は，異なる言語間の表現に適用すれば翻訳の技術となり，同一の言語内の表現に適用すれば言い換えの技術となる．機械翻訳への適用例を図1に示す．

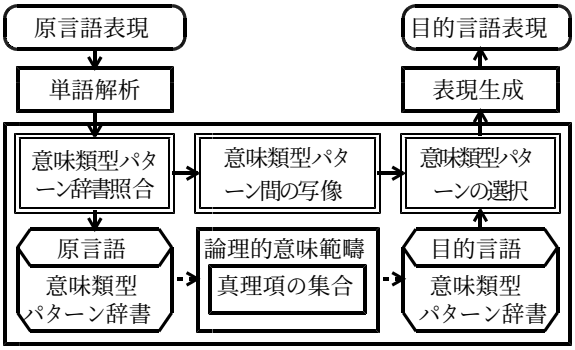

図1．「意味的等価変換方式」の構成

翻訳処理の手順は以下の通りである．

(1) 入力文と意味類型パターン辞書を照合し，適合する意味類型パターンを抽出する．

(2) 抽出された意味類型パターンに付与された「真理項」を介して，意味的に対応する目的言語の意味類型パターン（1つ以上）を発見する．

(3) 発見した意味類型パターンから最適なものを選択し，それを用いて目的言語の表現を生成する．

本方式は，あらかじめ対となったパターンを使用する従来の文型パターン方式と比べて，(2)の過程で意味的類似度の計算により複数の目的言語表現を発見しようとする点に特徴がある．

3．言語表現の構造と意味類型パターン

「意味類型パターン」は「文型パターン」を意味類型化したものである．文型パターンに意味コード（真理項）を付与し，付与された意味コードに基づいて意味分類する方法で作成されるから，「意味類型パターン辞書」を作成するには，あらかじめ「文型パターン辞書」を作成する必要がある．本研究では，文型パターン辞書を構築したので，以下では，その方法と結果について述べる．

(1) 言語表現の線形性と非線形性

文型パターンは，基本的には従来のパターン翻訳で使用されるテンプレートと同類のものである．本研究では，認知論的な立場からパターン化すべき対象表現とパターン化の方法を明確にした．

さて，文型パターンは，意味類型化されていないとはいえ，意味類型パターンと同様，対象に対する話者の認識を表現するための枠組みである．通常，話者が自分の考えを表現しようとするとき，思い浮かんださまざまな表現構造の中から適切なものを選び，全体の意味が損なわれないように注意深く構成要素を選択して表現を完成させる．構成要素には，他の要素に置き換えると全体の意味が損なわれるものと，他の代替要素に置き換えても全体の意味は損なわれないような要素の2種類が存在するから，これを見極めることが大切である．

そこで，前者を非線形要素，後者を線形要素として区別し，用例文を対象に線形要素を汎化することによって文型パターンを作成する．

「表現要素」と「表現」の「線形性」，「非線形性」は以下のように定義する．

【言語表現の線形性と非線形性の定義】

一つ以上の代替要素が存在し，その要素に置き換えても表現全体の意味が変化しないような要素を「線形要素」と定義する．次に，このような「線形要素」のみから構成される言語表現を「線形な表現」，1つ以上の非線形要素を有する言語表現は「非線形な表現」と定義する．

このように定義すると，線形表現では従来と同様の要素合成法が適用できるから，文型パターン化が必要となるのは，「非線形な表現」である．

(2) 意味の定義と線形要素の判定法

ところで，上記の定義を実際の表現に適用するには「表現の意味」とは何であるかを明確にする必要がある．意味の定義については言語学的にもさまざまな説が存在し，定説が存在しない．しかし，計算機から見れば，どのような定義の仕方も記号に過ぎず，排他的識別が可能な体系であれば問題はない．

そこで，本研究では原言語（日本語）表現の意味を目的言語（英語）の表現で定義することにする．この方法は，実際の対訳表現に対して，線形要素を比較的容易に判定することができ，対応する文型パターンから直接目的言語の表現構造が得られるから，機械翻訳のシステムを開発する上で大変便利である【付録1参照】．

図2に対訳文の例を示す．日本語原文は，「誰が」「どこを」卒業するかなどの詳細な内容を持っているが，それらの細かい点は無視して，「何かの事象の直後，誰かが何かの行為をする」と言う「事象間の関係」を表す日本語の表現だと考える．また，英語側では，そのように抽象化された意味（上位概念）が英文で定義されていると考える．

このように考えると，日本文中の名詞「彼女」，「大学」，名詞句「地元の小さな会社」などには，英文中に対応する要素が存在すること，また，これら部分には代替可能な要素が存在し，それを置き換えても英語構造は変化しないことが分かる．すなわち，事象間の関係として特定された意味は変化しないから，これらの部分は線形要素だと判定される．

(3) 線形要素の重要な性質

以上の定義により，言語表現から文型パターンを作成する際の判断基準として以下の指針が得られる．

①言語表現の意味を別の言語表現で定義している

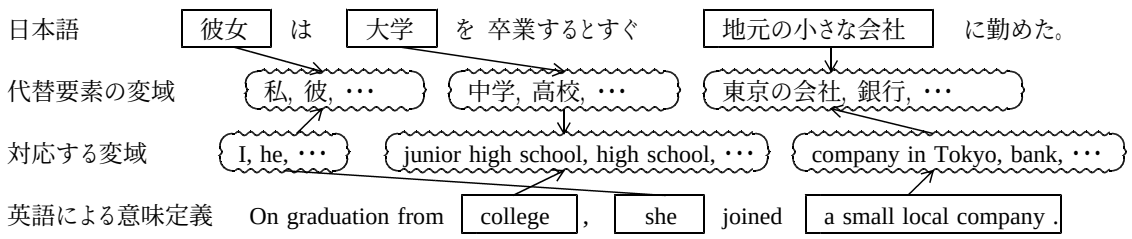

図2．線形要素の例

ため，線形要素の数や範囲は言語ペアに依存する．同族言語の場合は線形部分が増大し，異なる言語族間では減少することが予想されるが，これは翻訳の難易性を反映している．

②線形要素は置換可能だと言っても，何に置き換えても良い訳ではない．置き換え可能な範囲は，文法的，意味的に制限されるから，文型パターンではそれが明示される必要がある．

③線形要素のみからなる表現が線形な表現であるが，要素の範囲は任意に決められるから，表現全体の線形性，非線形性は要素の選び方に依存する．従って，汎用的な文型パターンを得るには，線形要素が多くなるように工夫すればよい．

④線形な要素だと言っても，それは表現全体から見たときの性質である．その要素自身が線形な表現だとは限らない．

(4) 線形性非線形性に着目した言語モデル

上記の性質（特に④）は，本研究の言語モデルとして大変重要である．図3に，本モデルによる言語表現の解釈例を示す．

<原文の表現>

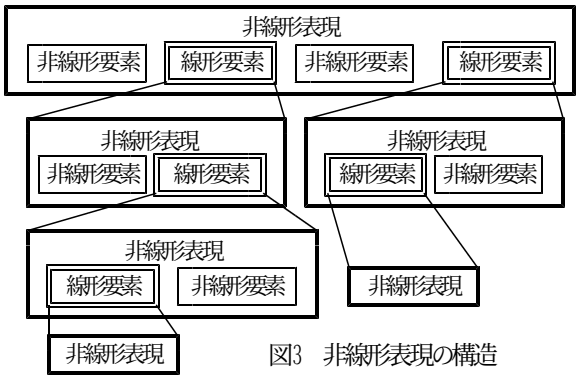

図3　非線形表現の構造

定義によれば，一般に言語表現は，この図のように非線形要素と線形要素から構成される．図3の原文は2つの線形要素と2つの非線形要素で構成される．

このうち，線形要素は意味的な代替要素が存在するから，意味の纏まる単位であり，それ自身は表現でもあるが，④で述べたように，線形表現だとは限らない．図の例では，線形要素は，いずれも非線形要素と線形要素からなる非線形表現である．

このように，非線形表現に含まれる線形要素を取り出していくと，最終的には線形要素を持たない非線形表現に帰着する．帰着した非線形表現は，単一の単語の場合もあるが，置き換え可能要素を持たない慣用句のような場合もある．

以上から分かるように，一般に言語表現は，最終的に1つ以上の非線形表現と0個以上の非線形要素に分けることができる．

この言語モデルで大切なのは，分解の各段階で出現する非線形表現が構造的に意味のまとまる単位だと言うことである．従って，要素分解によって元の意味を失わないようにするには，各段階の非線形表現に対して意味をすくい取る仕組みを持てば良いことになる．すなわち，文型パターンは，抽出可能な線形要素の表現のレベル（例えば，単語，句，節）に応じて作成すればよいことになる．

最近のフィルモアのConstruction　Grammar[24]や認知言語学のスキーマ[25]も表現構造の持つ意味に着目しているが，これらは線形要素間の関係に着目している．これと異なり，本研究の言語モデルは，非線形要素に着目している点が重要である【付録2参照】．

4. 文型記述言語

意味類型論に基づいた等価的変換方式を実現するには，従来の文型パターン辞書の規模を遥かに超える大規模なパターン辞書を開発する必要があり，そのような大量の文型パターン間の統一性と整合性を実現するには，記述能力に優れた言語が必要である．

本研究では，①汎化の程度に応じた記憶能力を持つこと，②文型パターン間で意味的に高い排他性が得られること，③大規模な対訳コーパスから半自動的に文型パターンが記述できることを狙って，文型記述言語を開発した．

表1．文型記述記号の構成

#	要素	種類用途	要素種別	
			非線形	線形
1	字面	日本語文字，英語文字	○	
2	変数 (15種類)	①単語変数(9種類)，②句変数(5種類)，③節変数(1種類)いも意味的な制約条件を持つ		○
3	関数 (107＋α種類)	①語形関数，②時制，相，様相関数，③品詞変換関数，④文型生成関数，⑤その他	○	△
4	記号 (7種類)	①表記の揺らぎ吸収，②任意要素の指定，③語順任意指定措，④位置変更可能指定，⑤省略要素補完指定，⑥その他	○	○

文型記述言語は，表1で示すように字面，変数，関数，記号の4種類の要素から構成される．字面が非線形要素の記述に使用されるのに対して，変数は線形要素（ただし，客体的表現）の記述に使用される．関数は，主として自制，相，様相など主体的表現の記述で使用され，非線形要素（字面の代わり）に使用される関数と線形要素指定に使用される関数がある．また，一部，表記の揺らぎなど（これも線形要素である）を吸収するための関数も存在する．記号は，わざわざ文型パターンに記述しなくても良いような任意な要素（線形要

素）を表すもの，語順を変えても意味は変わらないような要素を表すものなどさまざまである．

5．文型パターン辞書の構築

(1) 標本データと文型パターン化

日本語表現のうち用言と格要素から構成された単文の非線形構造については，既に「日本語語彙大系」を開発し，高品質の翻訳(正解率90%)が実現されている [26]．本研究では，重文と複文を対象に文型パターンを作成した．但し，述部を2つまたは3つ持つ文に限定した．これは，それ以上の長い文は，ある程度短い文に分解して扱うことができると考えたためである．

具体的には，まず，さまざまな対訳文書約30種類から100万文の対訳コーパスを作成し，対象となる標本文15万件を抽出した．得られた対訳文を対象に，①形態素解析，②線形要素の抽出，③線形要素の汎化を行い，以下に示す3レベルの文型パターンを作成した．

＜単語レベル＞線形な自立語を単語変数化したもの

＜句レベル＞単語レベルの文型パターンに対して，線形な句を変数化したもの．但し，単語変数を句変数化する場合もある．

＜節レベル＞句レベルの文型パターンに対して，線形な節を節変数化したもの．

作成した文型パターンの例を表2に示す．

(2) 作成された文型パターンとその特徴

各レベルの汎化において変数化された要素数と最終的に得られた異なり文型パターン数を表3に示す．

汎化された線形要素の割合を見ると，単語レベルでは，自立語約76万語のうち約47万語が単語変数化された．線形な自立語の割合は62%である．同様，名詞句，動詞句などの句のうち線形なものは22%である．これに対して，変数化できた節はきわめて少なく4.3%であった．

非線形要素は，それを取りだして翻訳し，元の文に組み込んでも意味的に適切な翻訳結果は得られない．上記の結果から，特に重文や複文では，複数の単文に分けて翻訳し，後で結合する方法（要素合成法）では，対訳例文に示されるような質の良い翻訳はできないことがわかる[*1]．

次に，異なり文型パターン数は，全体で21.5万件であり，単語レベル，句レベルの順に減少している．特に節レベルの文型パターンは単語レベルの約1/10で大変少ない．これは，線形な節が少なかったためである．

6．文型パターン辞書の評価

(1) 実験の条件と評価の方法

入力文との文型照合実験によって，文型パターン辞書の被覆率特性を評価した．実験の条件は以下の通りである．

①変数に付与された意味的な制約条件は無視する

変数の意味的制約条件の使用方法と効果は、適合パターンの絞り込み方式の中で検討中であ

表2．作成された文型パターンの例

レベル	言語	文型パターン	例　文
単語レベル	日本語	#＜N1は＞／V2て／N3を／N4に／V5.tekita。	うっかりして定期券を家に忘れてきた。
	英語	I was so AJ(V2) as to V5 #[N1_poss] N3 at N4.	I was so careless as to leave my season ticket at home.
句レベル	日本語	NP1は／V2.ta／N3に／V4.teiruのだから／N5.dantei	その結論は誤った前提に基づいているのだから誤りである。
	英語	NP1 is AJ(N5) in that it V4 on AJ (V2) N3.	The conclusion is wrong in that it is based on fales premise.
節レベル	日本語	CL1.tearuので、N2に当たっては／VP3.gimu。	それは極めて有毒であるので、使用に当たっては十二分に注意しなくてはならない。
	英語	so+that(CL1, VP3.must.passive with subj(CL1)_poss N2)	It is fignificantly toxic so that great caution must be taken with its use.

表3．線形要素の汎化と異なり文型パターン

文型パターンのレベル	線形要素の汎化			作成された異なり文型パターン数			
	全要素数	変数の数	線形な割合	重文	複分	複重文	合計
単語レベル	763,968	472,521	62%	59,658	49,897	12,174	121, 729
句レベル	463,636	102,000	22%	42,308	36,016	10,026	88,349
節レベル	267,601	11,486	4.3%	5,938	3,996	1,551	11,485
合計	－－	－－	－－	107,905	89,909	12,750	221,563

*1　文型パターンの元となった対訳例文では、日本文が重文と複文であるのに対して、英語訳文の多くは単文となっている。また、単文化する方法は、実にさまざまである。

り，別途報告する．

②実験はクロスヴァリデーションの方法で行う

具体的には，テスト用の入力文には，文型パターン作成用の例文の中からランダムに1万文を抽出して使用する．但し，各入力文がその入力文から作成された文型パターン（自己パターンという）へは必ず適合するため無視し，自己パターン以外への適合のみを評価する．

ところで，入力文に対して適合する文型パターンは通常複数存在し，それらがすべて意味的に正しいとは限らない．そこで，被覆率は以下の3つのパラメータによって評価する．

1）適合率（R）：入力した文のうち，1文型パターン以上に適合した入力文の割合

2）正解率（P1）：適合した文型パターンが意味的に正しい割合

3）累積正解率（P2）：1入力文に適合した文型パターンの中に意味的に正しい文型パターンが1つ以上存在する割合

このうち，Rは，①で述べたように意味的な適切性を無視しているから構文的な被覆率と言える． P1は適合した文型パターンからランダムに1つを選択するときの正解率である．また，P2は正解候補選択が上手く行ったときの最大の正解率を意味する．なお，最終的な意味的被覆率は，R×P2で評価する．

(2) 適合率の飽和特性

単語レベル，句レベル，節レベルの文型パターンの数と適合率の関係を図4に示す．

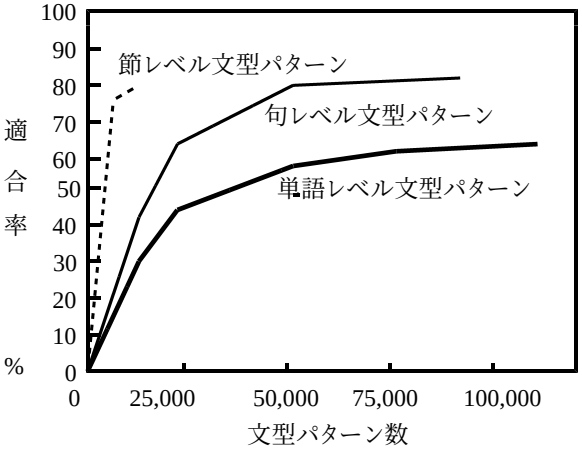

図4． 文型パターン適合率の飽和特性

この図から，適合率の飽和傾向が顕著である．横軸の文型パターンを使用頻度順に並べ替えると，飽和の速度は5倍程度速くなる．また，現状の文型パターンは様相，時制など，まだ汎化可能な余地を残しており，適合率はさらに向上する可能性がある．また，各レベルの文型を組み合わせて使用すれば，より高い被覆率が得られるとから，最終的には，合計10万件以下の文型パターンによって，かなり高い被覆率が得られる可能性がある．【付録3参照】

(3) 適合率と正解率

次に，文型パターン辞書全体の適合率と正解率を表4に示す．

表4．適合率と正解率

文型パターンのレベル	適合率 (R)	正解率 (P1)	累積正解率(P2)	意味的な被覆率(RxP2)
単語レベル	64.7 %	25 %	67 %	43.3 %
句レベル	80.0 %	29 %	69 %	55.2 %
節レベル	73.7 %	13 %	68%	50.1 %
合　計	91.8 %	－－	－－	70 %

この表では，適合率，意味的な被覆率共に句レベルの文型パターンが最大である．節レベルの文型パターン数は単語レベルの1/10しかないが，その割に高い被覆率を示している．パターン毎の汎用性の点から見ると，節レベルの文型パターンは，単語レベルと比べて10倍以上広い範囲をカバーすることになる．

(4) 意味的被服率

3種類の文型パターンは，単語レベル，句レベル，節レベルの順に，汎化されており適用範囲は広くなるが，逆に意味の曖昧さが増大する．このことから，複数のレベルの文型パターンに適合する入力文の場合は，この順に意味的に適切なものを選んで使用するのが良いと思われる．そこで，この順に文型パターンを選択して使用するときの意味的な被覆率を評価した．結果を図5に示す．

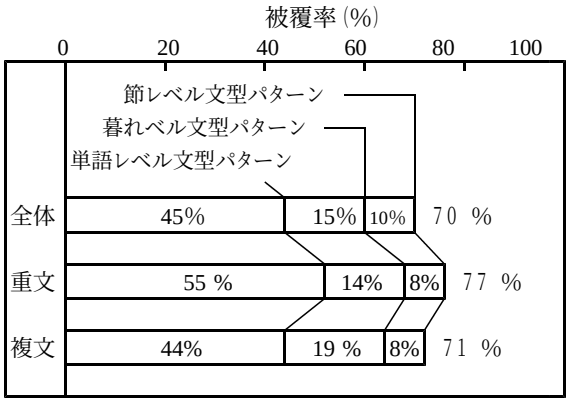

図5．文型パターン辞書全体での意味的な被覆率

図に示すように，3レベルの文型パターンを組み合わせて使用する場合の被覆率は70%であった．また，2節からなる重文，複文の意味的な被覆率は，それぞれ77%，71%であった．すでに述べたように文型パターンは，非線形な言語表現を対象にしたものである．線

形な表現の場合は要素合成法に基づいた従来の方法が適用できるので, 今後, 両者を併用した機械翻訳方式を実現すれば, 翻訳の品質は大幅に向上することが期待される.

7. まとめ

意味的等価変換方式の実現に必要な「意味類似パターン辞書」の構築をめざし, 本研究では, 重文と複文を対象に, その基礎となる「文型パターン辞書」の第1版を開発した. 開発した文型パターンは, 単語レベル (122,642件), 句レベル (80,130件) 節レベル (12,450件) で, 合計 (215,222件) である.

言語表現と表現要素の線形性, 非線形性に着目して対訳用例を半自動的に汎化することにより, 従来にない規模の文型パターンを開発することができた.

生成された文型パターンでは, 線形要素として変数化された自立語は60%, 句は22%であったのに対して, 節は4.3%に過ぎなかった. このことから, 重文, 複文の翻訳において, 従来のように節毎に訳してその結果を組み合わせて訳文を生成する方法では, 用例にあるような品質の良い訳文は得られないことが分かった. また, 句表現の中にも, 分離して単独で訳せないようなものが多く, これらを非線形な表現としてパターン化することの重要性が感じられる.

また, クロスヴァリデーション法による評価実験では, 文型パターン辞書全体での意味的な被覆率は71-77%であった. 第1段階としては, まずまずの被覆率が得られたと思われる.

ところで, この辞書はまだ時制様相などの部分で, されに汎化可能な部分を残している. 今後は, これらの汎化を行う一方, 平行して意味的な類似化を行い, 意味類型文型パターン辞書として完成させたい. その後は, 等価的変換方式による機械翻訳システムに適用し, その効果を明らかにする予定である.

謝辞

この研究は, 日本科学技術振興機構 (JST) の戦略的基礎研究推進事業 (CREST) の支援によって行われたものである. 関係各位および研究グループの方々のご協力に深謝する.

参考文献

[1] 長尾眞:自然言語処理, 岩波書店, 1996.
[2] M, Nagao: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, In A. Eithorn and R. Barneji (Eds.), Artificial and Human Intelligence, North-Holland, pp.173-180, 1984.
[3] Satoshi Sato: An example based translation and system, Proceedings of COLING-91, pp. 1259-1263, 1992.
[4] 佐藤理史:アナロジーによる機械翻訳, 共立出版, 1997.
[5] P. F. Brown, C. John, S. D. Pietra, F. Jelinek, J. D. Lfferty, R. L. Mercar and P. S. Roossin: A Statistical Approach to Machin Translation, Computational Linguistics, Vol. 16, No. 2, pp. 79-85, 1990
[6] T. Watanabe and E. Sumita: Bidirectional Decoding for Statistical Machine Translation, Proceedings of COLING-02, pp. 1075-1085, 2002
[7] K. Takeda: Translation, 34th Annual Meeting of the Association for Computational Linguistics, pp. 144-151, 1996.
[8] K. Takeda: Pattern-based Machine Translation, the 16th COLING, Vol. 2, pp. 1155-1158, 1996.
[9] H. Watanabe and K. Takeda: A Pattern-based machine translation system extended by example based processing, 17th COLING, pp.1369-1373, 1998.
[10] 内野一, 白井諭, 横尾昭男, 大山芳史, 古瀬蔵:速報型日英翻訳システムALTFLASH, 電子情報通信学会論文誌, Vol.J84-D-II, No.6, pp.1168-117, 2001.
[11] 田中穂積監修:「自然言語処理-基礎と応用」電子情報通信学会, 岩波書店, 1998
[12] 長尾眞, 黒岩禎夫, 佐藤理史, 池原悟, 中野洋:「言語情報処理」, 岩波書店, 1998
[13] F. Almuallin, Y. Akida, T. Yamazaki, A.Yokoo and S. Kaneda: Two Methods for Leaning ALT-J/E Translation Rules from Examples and a Semantic Hierarchy, COLING94, pp. 58-63, 1994
[14] H. A. Guvenir and Cicekli: Leaning Tanslation Rules from Examples. Information systems Vol. 23. No. 6, pp. 2325-363, 1988
[15] M. Kitamura and Y. Matsumoto: Automatic EXtraction of Word Sequence Correspondence in Parallel Corpora, 4th Annual Workshop on Very Large Corpora, pp. 79-87, 1996
[16]池原悟, 宮崎正弘, 白井諭, 林良彦: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, Vol. 28, No. 12, pp. 1269-1279, 1987
[17] 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:「日本語語彙大系」岩波書店 1997.
[18] S. Shirai, S.Ikehara, A. Yokoo and H. Inoue: The Quantity of Valency Pttern Pairs required for Japanese to English MT and Their Compilation, Proc. of NLPRS'95, Vol. 1, pp. 443-448, 1995
[19] 池原諭:自然言語処理の基本問題への挑戦, 人工知能学会誌, Vol.16,No.3, pp. 522-430, 2001.
[20] S. Ikehara: Meanig Comprehension Using Semantic Patterns in a Large Scale Knowledge-Base, Proceedings of the PACLING'01, pp. 26-35, 2001
[21] 池原悟: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, No.33, pp.1-7, 2002.
[22] 有田潤:「ドイツ語講座II」南江堂, pp. 48-56, 1987.
[23] 市川亀久彌 「独創的研究の方法論」(増補版), 三和書房, 1963
[24] C. Fillmore, P. Kay, L. Michaelis and I. Sag: Construction Grammar, Stanford Univ Center for the Study, 2005.
[25] R. W. Langacker:FOUNDATION OF COGNITIVE GRAMMAR, Stanford University Press, 1987.
[26] 金出地真人, 池原悟, 村上仁一: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, 6Y-04, 2-267-268, 2001.
[27] 三浦つとむ:「認識と言語の理論」第1部～第3部, 勁草書房, 1967

【付録1】意味を定義する方法の問題

たとえば，関係意味論（三浦つとむ）[27]では，言語表現の意味は，話者の対象認識と表現との対応関係だとされているように，厳密に言って，言語表現の意味は表現ごとに固有のものであり，当該表現以外の表現で正しく表すことは不可能である．従って，異なる言語表現の意味が同一であることはあり得ず，ましてや，他の記号や別の言語を用いてそれを厳密に定義することは不可能である．

言語処理では，このような厳密な意味での意味の同等性は問題とせず，近似としての意味の精度を問題としている．言語表現に結びつけられた話者の認識が誤解されない範囲で正しく扱えるかどうかが問題である．（この点では，言語表現自体も話者の認識が厳密に正しく表現できているとは言えず，近似でしかない．）

ここで，文型パターン化で問題となるのは，抽象化された表現構造の持つ規範としての意味を何によって表現するかの問題であり，本研究では，これを英語表現で定義することとした．

ところで，言語表現の意味を別の言語表現で定義する際に問題となるのは，一般に言語表現には異型同内容の表現と同型異内容の表現が存在するため，元の表現の意味がユニークには定義できない可能性があることである．しかし，以下の理由で，本研究の目的は損なわれないと判断できる．

まず，異型同内容の表現は，形式が異なるとは言え，いずれも元の表現の意味を表すから，言語表現の意味的な等価変換において問題とならない．むしろ，「複数の目的言語表現から文脈に合った訳文構造を選択して訳せるようにしたい」と言う本研究の目的の一つからみて，好都合である．

次に，同型異内容の表現の場合は，意味定義に使用された表現が，原文と異なる意味を併せ持つことになるが，元々計算機自身が意味を理解しているわけではない．機械翻訳や言い換えでは，計算機は人間が理解できる結果を出力すればよい．人間が，多義のある表現を聞いたり読んだりするのは日常のことであり，そのような結果が出力されても，人間にどちらの意味が正しいかが分かればそれでよい．

【付録2】非線形要素と文型パターンの関係

映画や絵画などさまざまな表現の持つ意味について，全体の意味が先か，部分の意味が先かの問題は哲学的に重要な問題として議論されてきたが，言語表現においてもこの問題は大変重要である．従来の自然言語処理は，部分の意味から減退の意味が説明できることを仮定した要素合成法が基本となっているが，両者は，共に重要と考えられる．

そこで，本研究では，部分の意味を特定するのに先立って「全体の意味を先につかむ」方法を実現することを目指している．本研究で，文型パターン化を試行している理由は以下の通りである．

第1に，全体の意味をすくい取る仕組みとしては，線形要素ではなく，むしろ非線形要素に着目した方法が必要だと考えられること，第2に，非線形要素は代替不能であるから，字面で表記せざるを得ないこと，また，第3に，その出現位置も固定的であることを考えると，非線形な要素の配列を表現するにはパターンが適していると考えられることである．

なお，本研究では，重文と複文を対象としているが，単文が単独の事象を表す表現の枠組みであるのに対して，重文や複文は事象と事象との関係を表す枠組みである．従って，本研究の目的は，事象間の関係の意味を掬いとるため枠組みを実現することである．

【付録3】文型パターン数とプラトンの問題

言語表現の多様性を指摘したものに，プラトンの問題がある．これは，子供の言語能力の獲得に関する問題としてよく知られているが，「人間（個人）は，有限の記憶容量しか持たないのに，どのようにして無限とも言える多様な表現が使えるのか」と言い換えられる．

プラトンが指摘したのは，元々個人の表現能力の問題であるが，言語表現の多様性は社会的集団として獲得されたものである．また，歴史的に蓄積された知識量を考えると，社会的な集団の記憶容量もほぼ無限と言って良い．従って，言語表現の多様性は，個人の使用する範囲を超えており，言語処理では，これにいかに取り組むかが重要な問題である．

いままで，さまざまな言語モデルがあるが，言語処理は，なかなかこの問題に答えられないままである．この問題がいかなる方法で説明されるかによって，その言語モデルの先行きが占えそうである．

本研究の言語モデルは，「すべての言語表現は，有限の非線形な表現の組み合わせから構成される」と説明している．すなわち，有限の非線形表現が組み合わされることによって可能な表現は幾何級数的に増大し，ほぼ無限と言える言語表現が生成されることを主張している．

非線形な表現数が有限であるとする根拠は，①「非線形な表現は，その構造を覚えておかないと使えない」のに対して，②「記憶容量が有限の人間がこれを使いこなしている」ことによるものである．従って，本研究の言語モデルの正しさを示すには，文型パターン数と被覆率の関係を明らかにする必要がある．

この問題の結論はまだ出せないが，表現の無限性は，厳密な無限ではなく，数え上げることが技術的に困難であるという意味での無限であるから、言語モデルの妥当性を示すには，厳密に100％の被覆率が得られなくても，技術的な扱いが容易な範囲で文型パターン数で、高い被覆率を得ることが必要だと思われる．