第10回LACE研究会（2005.12.24-25）

非線形性に着目した言語表現モデルと重文と複文に対するパターン辞書の開発

Non-linearity based Linguistic Model and Pattern Dictionary Development for Complex and Conpound Sentences

池原 悟（鳥取大学）

あらまし

言語表現の非線形性に着目した言語表現モデルとそれに基づくパターン化の方法を提案し、日本語重文、複文に対する文型パターン辞書を開発した。人間の思考の枠組みとも言われる言語表現の形式をパターン化し、網羅的に収集することができれば、それは、言語表現の意味をすくい取る網の目として、機械翻訳だけでなく幅広い応用が期待される。本研究では、まず、約30種類のドキュメントから、基本的な文例100万件を取り出し、日英対訳コーパスを作成した。次に、その中から、２つまたは３つの述部を持つ重文、複文15万件を抽出し、それに含まれる線形な要素を半自動的に発見して汎化することにより、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）からなる文型パターン辞書（合計21.5万件）を作成した。その結果によれば、汎化することができた線形要素は，自立語で62%（47万語／76万語）であったのに対して、句では22%（10万件／46万件)、節では4%（1.1万件／27万件）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．この結果から，複文、重文の多くは、複数の単文に分けて翻訳し後で結合するなどの方法では、対訳例文に示されるような訳文は得られないことが分かった。また、クロスバリデーションによる文型パターン辞書の評価実験では、３種類のパターン辞書全体での構文的な被覆率は92%であるが、入力文が意味的に不適切な文型に適合することも多く、それを除いた意味的な被覆率は70%であった。基本とも言える述部が２の重文，複文の被覆率は、それぞれ77%，71%である．パターン辞書は，非線形な言語表現を対象にしている．線形な表現の場合は従来の要素合成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．また、パターンは、意味を掬くい取るための網の目である。機械翻訳だけでなく、広く意味解析への利用が期待される。

１．まえがき

機械翻訳では、パターン翻訳方式、トランスファー方式、用例翻訳方式などさまざまな翻訳方式が研究されてきた（長尾1996、池原1998、田中1998)。最近、統計翻訳方式（Brown et. al. 1990, Watanabe and Sumita 2002, Vogel et.al. 2003）が注目されているが、直訳の可能性の高い同一言語族間の翻訳に比べて、異なる言語族間での翻訳では、困難さが予想される。実用システムの多くは、依然としてトランスファー方式が基本である。この方式は、要素合成法に基く方法で、統合構造と意味を分離して翻訳するため、原文の意味が失われる点に問題がある。

これに対して、パターン翻訳方式（Takeda 1996, Watanabe and Takeda 1998）と用例翻訳方式(Nagao 1984、Sumita and Iida 1991, Sato 1992, Brown 1999）は、統語構造と意味を一体的に扱う方式であり、高品質の翻訳が期待できる。しかし、パターン翻訳方式では、あらかじめ大規模なパターン辞書を開発する必要があり、パターン間での意味的な排他性の保証が困難であることが問題であった。このため、トランスファー方式に特定分野の小規模なパターン辞書を組み合わせたハイブリッド型のシステム(Jung et. al. 1999, 内野ほか 2001)が多い。

これに対して用例翻訳方式は、構造的に一致する対訳用例の中の意味的に類似する単語や表現の一部を置き換えて訳文を得る方法であり、あらかじめパターン辞書を準備する必要はない。しかし、用例中の要素の置換の可能性は、用例毎に異なるため、その判定を自動化することが難しい。この問題を解決するには、各用例に対してあらかじめ置換可能な要素を指定しておけばよいが、その方法は、パターン翻訳に帰着する。

この問題を解決する方法として、認知言語学（Langacker 1987）のCG(Cognitive Grammar)やフィルモア（Fillmore 1988, Fillmore et.al. 1988, Fillmore et.al. 2005）の構文文法（CxG：Construction Grammar）の方法が注目される。これらの方法では、部分の意味と全体の意味を関連づける方法についてさまざまな工夫が行われている。しかし、意味の定義が明確でないため、構造的な意味の単位を決定する基準が曖昧である。

ところで、表現構造と意味を一体化した方法（Ikehara 2001）としては、すでに、日英機械翻訳システムALT-J/E（池原ほか1987）において、単文構造を対象とした表現意味辞書「日本語語彙体系」（池原ほか1997）が開発されている。この辞書は、動詞と

格要素との関係を17,000件の結合価パターンにまとめたものである。格要素となる名詞の意味的な用法は、名詞意味属性（2700分類）を用いて指定されており、40万語の名詞意味辞書と連動する。翻訳実験では、IPAL（IPA　1987）の例文5000件に対して、90％の正解率が得られており、方式限界は97%であることが報告されている（金出地ほか2001)。

これにより、日英機械翻訳における単文の意味的な訳し分けの問題はほぼ解決できたとみられるので、本研究では、重文と複文を対象とした文型パターン辞書を開発する。

ところで、今まで対訳用例からパターンを収集する方法について、人工知能の分野では、ID4の学習アルゴリズムを応用したフレーム獲得などの研究（Almuallin et.al. 1994, Guvenir 1988）が行われ、言語処理の分野でも、多くの研究（Kaji et.al. 1992, Rillof 1996, Kitamura and Matsumoto 1996, Watanabe et.al. 2000, Yangarber 2000）が行われている。しかし、統計的な手法を基本とするこれらの学習方法で得られるパターンは、頻度の高い表現に限定されており、網羅的で実用的な品質のパターン辞書を作成することは難しい。

「日本語語彙体系」の開発では、適切な計算機環境を準備すれば、言語アナリストは、頻度が低い用例からでもその背後に存在する規則をパターンとして能率良く取り出せることが知られている(Shiai et.al.1995、池原2001)。問題は、いかにして人手作業を効率化する仕組みを実現するかである。

本研究では、言語表現の非線形性に着目した言語表現モデルとそれに基づくパターン化の方法を提案し、パターン辞書開発の効率化を図った。これは、対訳用例から意味のまとまる非線形な言語表現を取り出し、内包される線形要素を汎化してパターンを得る方法である。

2．言語表現のモデルとパターン化

「人間は対象を概念化する過程である種のフレームワークを使用していること」、また、「そのフレームワークとしては話者の母国言語の表現の枠組みが用いられること」が指摘されている(有井1987)。

パターン辞書は、このような表現のフレームワークをパターンとして網羅的に収集することを狙ったものである。与えられた言語表現の意味をすくい取る網の目として、機械翻訳だけでなく、さまざまな自然言語処理に利用できる可能性がある。

2．1　従来のパターンの問題点

すでに、網羅性の高い文型パターン辞書として、単文を対象とした（「日本語語彙大系」）が実現されているが、単文全体の構造に対しては必ずしも適切な訳文構造が得られない場合がある。その原因は、単文の構造と意味を結合価パターンで記述したためと考えられる。

結合価パターンは、動詞と名詞を中心とする自立語の意味的な関係構造を表現したものである。付属語（助詞、助動詞など）の意味は分離して扱われるため、元の意味が失われることがある。この問題を解決するには、構文全体の中で付属語の意味を捉える仕組みが必要である。

これに対して、構文文法では、自立語と付属語を含む表現全体の意味を扱うための仕組みが考えられているが、パターン化すべき表現の範囲と汎化可能な要素の範囲を決定するための基準が明確でない。

一般的に、パターン化と汎化の対象範囲は意味のまとまる範囲であるが、機械翻訳では、翻訳対象となる言語のペアによって、要求される意味の分解能は大きく異なるから、言語ペアに応じて意味のまとまる範囲も異なる。

以上の２つの問題を解決するため、本章では、(1)「非線形性に着目した言語表現モデル」と定義する。また、(2)「表現の意味を他の自然言語表現を使用して記述すること」によって、パターン化と汎化の対象範囲に対する判定基準を明確にする。

2．2　言語表現の線形性と非線形性

筆者が思い浮かべた表現のフレームワークには、他の要素に置き換えると全体の意味が損なわれる要素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。

そこで，前者を「非線形要素」，後者を「線形要素」と区別し，「表現要素」と「表現」に対する「線形性」，「非線形性」を以下のように定義する．

＜定義１＞線形要素と非線形要素

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」、それ以外の要素を「非線形要素」と定義する。

＜定義２＞線形表現と非線形表現

「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。

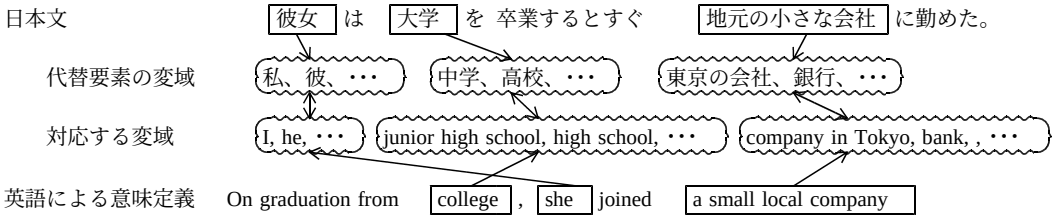

図１．線形要素の例

但し、定義１の「表現全体の意味」は、表現構造の表す意味（「抽象化された複合概念」[*1]）である。本研究では、当該言語とは異なる任意の言語の表現を用いて定義する[*2]。

図１に日英対訳文の例を示す。原文は、「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」（「抽象化された複合概念」）を表す日本語の表現で、その意味は英語表現で定義されている。個別概念である「彼女」、「大学」などには、代替可能な要素が存在し、それを置き換えても「抽象化された複合概念」（英語表現構造）は変化しないから、これらは、線形要素である。

２．３　非線形要素の基本的特徴

上記の定義では、線形要素は以下の４つの重要な特徴を持つことが指摘できる。これらの特徴から、パターン化のための重要な指針が得られる。

＜特徴１＞線形要素の言語ペア依存性

言語表現の意味を別の言語表現で定義しているため、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要素の範囲が増大し、異なる言語族間では減少することが予想されるが、これは言語による翻訳の難易性の違いを反映している。

＜特徴２＞代替要素の有限性

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、パターンではそれが「変域」として明示される必要がある。

＜特徴３＞線形要素と非線形要素の相対性

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。

＜特徴４＞線形要素と非戦役表現の同時性

「線形要素」が線形であるのは、あくまで表現全体から見たときの話であり、それ自身は非線形な表現であっても良い。

２．４　言語表現モデル

前項で示した特徴から導かれる言語表現モデルについて述べる。定義１によれば、言語表現は「非線形要素」と「線形要素」から構成されるが、特徴３によれば表現要素の範囲は任意に選択できるから、意味のまとまる範囲の表現（例えば、単語、句、節など）の中から「線形要素」を抽出することとする。このようにして抽出した「線形要素」は、特徴４により、それ自身「非線形表現」でもよいから、言語表現は、一般に図２のような言語表現モデルで表すことができる。

この図から分かるように、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、置き換え可能な要素を持たない慣用句のような場合もある。

以上から、本研究の言語表現モデルでは、言語表現は、１つ以上の非線形表現と０個以上の非線形要素から構成される。

２．５　非線形な表現を記述する方法

図２の言語表現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」が意味のまとまる表現の単位だと言うことである．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する意味辞書を持てばよい。例えば、言語表現を文、節、句、単語のレベルに分類し、その中の非線形な表現を対象とした意味辞書[*3]を構築すれば、文全体の意味をすくい取るための仕組みは一通り揃うことになる。

ところで、表現の構造を表現する枠組みとしては、認知言語学や構文文法などの方法もあるが、

(1) 非線形要素は通常字面表記が適していること
(2) 線形要素と非線形要素の出現順序は固定的な場合が多く任意性が少ないこと

を考えると、パターンが適している。そこで、本研究では、意味のまとまる表現をパターン化する。

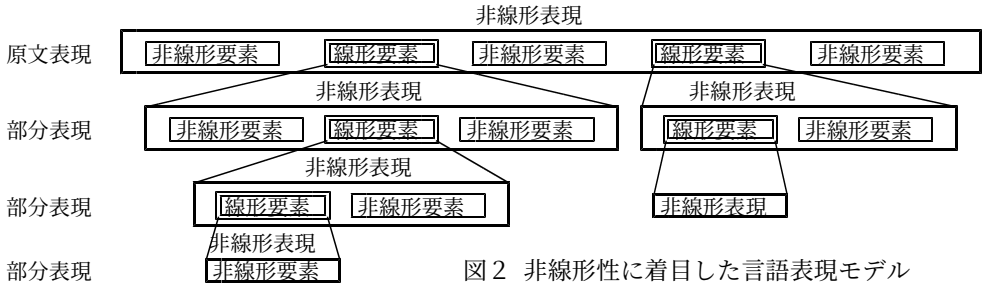

図２ 非線形性に着目した言語表現モデル

*1 ここでは、言語で表現される「概念」を「単一概念」（単語で表現可能な概念）と「複合概念」（複数の単語からなる表現で表わされる概念）に分類し、複数の複合概念間に共通した上位概念を「抽象化された複合概念」と言う（三浦1967、池原2003)。例えば、物事や事象を比較する表現の集合において、個々の表現の表す概念が「複合概念」であるのに対して、表現集合全体に共通する概念（「比較」）が「抽象化された複合概念」である。

*2 計算機では、どのような意味記述も単に記号にしかすぎないから、意味論的に矛盾のない記号体系で表現できればよい。本研究では、日英機械翻訳を考え、日本語表現の意味を英語表現で定義する。このように、言語表現の意味を他の自然言語で定義する場合、定義に使用した言語側での意味的な多義が問題となる。しかし、機械翻訳の場合は、翻訳結果の意味を理解するのは目的言語側の人間であるので、あまり問題にはならないと考えられる。

*3 単一の単語の場合は単語対訳辞書が相当するから、複合語のパターン辞書を持てばよい。

３．パターンの記述方法

原言語表現の意味を目的言語表現を用いて定義することについてはすでに述べた。本研究では、日英機械翻訳への適用を狙って、原言語を日本語とし、その表現の意味は英語で定義されるものとする。この典型的な例は、日英対訳文である。そこで、以下では、日英対訳例文を汎化することによってパターン対を生成することとする。

３．１　パターン化の原則

定義１から、パターンは線形要素と非線形要素の２種類の要素からなり、線形要素であるのは、以下の２つの場合である。

(1) 日本語側の表現要素に対して英語側に対応する要素がある場合
(2) 英語側にそれに対応する要素はないが、日本語側のその要素を削除しても対応する英語表現は変化しない場合

そこで、与えられた日英対訳表現を対象に、これらに該当する部分表現を抽出して汎化することによってパターンを生成する。

３．２ パターン記述の基本的枠組み

日英対訳パターンを記述するため、
(1)汎化の程度に応じた記述能力を持つこと，
(2)パターン間で意味的な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パターンが生成できること
を目標にパターン記述言語を設計した。パターンの記述に使用される要素は、表１に示すように字面、変数、関数、記号の４種類である。

表１．パターン記述言語

#	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) ･意味的な制約条件あり		○
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		○

表現要素のうち、自立語的な要素（単語、句、節）では、線形な要素は変数を使用して記述し、非線形な要素は、字面によって記述する。

これに対して、付属語的な要素（助詞、状動詞など）では、線形な要素は関数を用いて記述し、非線形な要素は字面または関数で記述する。

関数には、非線形な要素を指定するもの（字面の代わり使用される）と線形な要素を指定するものがあり、両者は区別して使用される。例えば、表現には、現在時制でしか使用できないもの（非線形）や過去形でも使用できるもの（線形）がある。このような、時制、相、様相などの表現の線形性、非線形性は、使用する関数によって識別される。

また、表記の揺らぎや語順の任意性など、特殊な線形要素は、記号を用いて記述する。

３．３　変数化する線形要素

（１）線形要素の変数化の原則

線形な単語、句、節を変数化する。要素の線形性を判定するための原則は以下の通りである。

①変数化する対象は、パターンから見たときの線形要素である。それ自身が非線形表現であっても良い。
②変数化する要素と対応する英語表現の要素が必ずしも同一の文法的属性を持つ必要はない。
③あってもなくても意味の変わらない要素（それも線形要素）は、記号で表示する。

（２）制約条件の付与

特徴２に基づき、日本語パターン側の変数に対して、意味的に置き換え可能な範囲を名詞意味属性、動詞意味属性、副詞意味属性などを用いて指定する。

３．４　任意化とグループ化

（１）必須要素と任意要素

パターン化では、「必須要素」と「任意要素」を明確に区別する。「必須要素」は、日本語パターン内にその要素がないと全体の意味が変わってしまい対応する英語パターンが決定できなくなるもの言う。これに対して、「任意要素」は日本語パターン内にその要素がなくても対応する英語パターンは変化しないものを言う。

このうち、「任意要素」は、さらに「原文任意要素」と「パターン任意要素」に分類する。「原文任意要素」は、日本語パターンでは、その位置と内容が指定されるが、英語表現側で指定する必要のないものである。また、「パターン任意要素」は、英語パターン側に対応する要素を指定する必要があるものである。例えば、訳語挿入位置など、英語側に指定がないと英語表現の生成が困難となるものである。

（２）表現要素のグループ化

線形要素には、表記上の揺らぎのように、変域が特定の表現に限定されるものが存在する。このような線形要素では、選択記号を使用し変域を具体的に指定する。

３．５　語順の扱い

言語表現では、語順や出現する位置を変更しても表現全体の意味は変わらないような要素が存在する。このような要素も線形要素である。以下の２種類の記号を使用して指定する。

（１）順序任意要素指定記号

格要素など語順を変更してもパターンの意味は変化しない要素をグループ化する。

（2）位置変更可能要素指定記号

副詞（副詞的表現を含む）など出現する位置が変わっても全体の意味が変わらない要素について、変更可能な位置を指定する。

3．6　文型パターンの例

文型パターンの記述例を表２に示す．以下、例に使用された変数、関数などの意味を簡単に説明する。

＜単語レベルのパターンの説明＞

・N1、N2、N3：「名詞変数」、V2、V5：「動詞変数」
・(G4)、(R3003)：変数に対する意味的制約条件を意味属性番号で指定
・#1[・・・]：省略しても良い要素（「パターン任意要素」）
・／：入力文にはなくても良い要素（「原文任意要素」）
・.tekita：述部語尾を指定するための関数（「語形指定関数」）
・AJ(V2)：動詞変数V2の値を形容詞化したもの（「変数変数」）
・N1_poss：N1の値を所有格に変形する（「語形指定関数」）

＜句レベルのパターンの説明＞

・NP1：「名詞句変数」、(他は単語レベルと同様)

＜節レベルのパターンの説明＞

・CL1：「節変数」
・so+that（・・・、・・・）：so　that構文を生成する「構文指定関数」
・subj(CL)：節変数の値から主語を抽出する「要素抽出関数」(他は単語レベルと同様)

5．文型パターン辞書の構築

２つまたは３つの述部を持つ基本的で標準的な重文、複文の対訳例文[*1]を用意し、それを汎化することによって文型パターン辞書を作成した。

但し、対応する英訳文の意味が単独で日本語文の意味に対応するものをパターン化の対象とし，前後の文脈から意訳されているなど，与えられた日本文だけでは対応関係を持たないような対訳用例は文型パターン化の対象としないこととした.

5．1　文型パターン作成の手順

（1）汎化のレベル

表現要素の線形性と非線形性を判断するには、あらかじめ表現要素の選び方を決める必要がある。そこで、言語表現の文法的な構成単位に着目して以下の３レベルの汎化を行った。

＜単語レベル＞：線形な自立語（名詞、動詞、形容詞、副詞など）を変数化したレベル。

＜句レベルの汎化＞：線形な句（名詞句、形容詞句、動詞句、副詞句など）を変数化したレベル。

＜節レベルの汎化＞：線形な節（連体節と連用節）を変数化したレベル。

ところで、「日本語語彙大系」では、単文は単一の事象を表現するための枠組だと考え、事象名を表す述部用言（動詞、形容詞）は変数化せず字面で記述されている。これに対して、重文と複文は、因果関係など事象間の関係を表現するための枠組みである。関係を表す接続詞、接続助詞などは字面で記述するが、述部用言は変数化の対象とした[*2]。

（2）汎化の手順

以下の手順で文型パターン辞書を作成した。

＜第1ステップ＞日英対訳コーパスの作成

日英・英日辞書、ハンドブック、日英機械翻訳試験文集など、約30種類のドキュメントから、日英対訳例文100万件を収集し対訳コーパスを作成した。

＜第２ステップ＞対訳例文の準備

上記のコーパスの日本文を形態素解析プログラム

表３．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(G4)は＞／V2(R3003)て／N3(G932)を／N4(G447)に／V5(R1809).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	NP(G1022)1は／V2(R1513).ta／N3(G2449)に／V4(R9100).teiruのだから／N5(N1453).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(G2492).tearuので、N2(G2005)に当たっては／VP3(R3901).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 significantly toxic so that great caution must be taken with its use.

*1 分野に固有の表現は、比較的その数も少なく、作成が容易であるので、本研究では、汎用性の高い文型パターンを網羅的に収集することを目指している。重文と複文を対象としたのは、すでに、単文では、「日本語語彙大系」が実現されているためである。また、述部数を２と３に限定したのは、述部数４以上の重文、複文は、述部数を３以下の文に分解して訳せる場合が多いと考えたためである。

*2 日本語で通常使用される用言は6000語程度であるから、単文パターンではこれを変数化しなくてもパターン数はその数倍程度に収まると見られる。これに対して複文重文では、汎化しない場合、パターン数は、6,000の２乗～３乗のオーダになり、汎用性の乏しいパターン辞書になるおそれがあるため、汎化は必須と考えられる。

表３　文型パターンの種類と異なりパターン数

文種別	説　明	作成した文型パターン数（）：重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	文接続１カ所を持つ文	53,313	33,898	5,930	93,141
文種別２	文接続２カ所を持つ文	5,622	3,270	309	9,201
文種別３	埋込み文１つを持つ文	45,678	30,756	3,923	80,357
文種別４	埋込み文２つを持つ文	5,591	4,024	772	10,387
文種別５	文接続と埋込文各１つを持つ文	12,438	8,182	1,516	22,136
－－		122,642	80,130	12,450	215,222

ALT-JAWS(ALT　2002)を使用して形態素解析し、該当する対訳標本文15万件を抽出した。また、その中に含まれる解析誤りを人手によって修正した。

＜第３ステップ＞意味的対応づけと変数化

上記の結果と日英対訳辞書を使用して，日英例文中の要素間の意味的対応関係を抽出し、変数化した。

＜第４ステップ＞関数化対象表現の抽出

ステップ２の結果を対象に、関数化の対象となる字面を抽出して関数化した。

＜第５ステップ＞その他の各種要素の記述支援

任意要素，補完要素，冠詞と丁寧表現など、記号表記の対象要素を抽出して、記号化した。

以上の過程で、線形要素であることが自動的に判定できるものについては，機械的な置き換えを行い、自動的判定の困難なものは、言語アナリストの判断にゆだねた。

なお、パターン化の対象となった対訳例文の平均単語数は、日本文が12.9語／文（最大63個）、英文が、10.3語（最大59語）である。

５．２　生成された文型パターン数

（１）生成された文型パターン数

得られた文型パターンの種類と異なりパターン数を表３に示す。単語レベルのパターンでは，123万件中、字面のみの文型パターンが642件、逆に字面を含まないものがxxx件存在する、次に，句レベルのパターン化では，単語レベルで得られた文型パターンのうちの約82%がさらに汎化され，8.0万件の文型パターンが得られた．

これに対して，節レベルで作成された文型パターンは1.2万件で，単語レベルに比べて約1/10である．これは大半の対訳例文の節は非線形要素であり，汎化困難であることを示している．

（２）汎化された要素の割合

各レベルの汎化において変数化された要素数を表４に示す．汎化された線形要素の割合は、単語レベルでは，62%（47万語／76万語）、句では、22%であったのの対して、、節ではきわめて少なく4.3%であった．

非線形要素は，それを取り出して翻訳し，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，従来のような要素合成法では，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]．

表４．線形要素の割合

要素種別	全要素数	変数の数	線形な割合
単語(自立語)	763,968	472,521	62%
句	463,636	102,000	22%
節	267,601	11,486	4.3%

６．文型パターン辞書の評価

６．１　実験の条件と評価の方法

入力文とのパターン照合実験によって，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変数の意味的な制約条件は無視する

入力文に適合する文型パターンの中から意味的に適切なパターンを選択する方法は別の課題と考え、ここでは、得られた文型パターンの潜在的な能力として被覆率を評価する。

②実験はクロスヴァリデーションの方法とする。

具体的には，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、入力文として使用する。但し，当該入力文から作成された文型パターンへは必ず適合するため無視し，それ以外のパターン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．

- **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
- **正解率（Ｐ１）：**適合した文型パターンが意味的に正しく、入力文の翻訳に使える割合
- **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く、入力文の翻訳に使えるパターンが１つ以上存在する割合
- **意味的被覆率（Ｃ）：**Ｒ×Ｐ２

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である。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

*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。

表５．文型パターン辞書の被覆率特性

文型パターンのレベル	平均適合文型数	適合率(R)	正解率(P1)	累積正解率(P2)	意味的被覆率(RxP2)
単語レベル	57 件	64.7 %	25 %	67 %	43.3 %
句レベル	470 件	80.0 %	29 %	69 %	55.2 %
節レベル	182 件	73.7 %	13 %	68%	50.1 %
合　計	－－	91.8 %	－－	－－	70 %

るときの正解率である．またＰ２は適合したパターンから正解候補選択が上手くできたときの最大の正解率を意味する．最終的な文型パターン辞書の被覆率は、意味味的被覆率Cで評価する。

６．２　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３に示す．

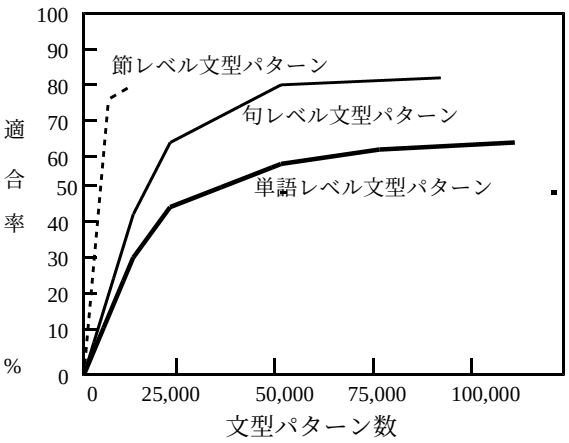

図３．文型パターン適合率の飽和特性

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．単文をほぼ網羅するために必要な結合価パターン数は、約2.5万件であると推定されている（Shirai et.al. 1995）。重文、複文の場合も、ほぼ網羅するために必要なパターン数も同じオーダに収まることが期待される。

６．３　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５に示す．この表から、文型パターン辞書全体では、構文上、入力文の90%以上をカバーすることが分かる。しかし、意味的に不適切で翻訳に使用できないパターンに適合する場合も多く、それらを除いたときの意味的な被覆率は70%である。

パターンの種別では、句レベルの文型パターンが適合率，意味的被覆率共に最大である．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，その割に被覆率は高い．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している．

６．５　意味的被覆率

３種類の文型パターンは，単語レベル，句レベル，節レベルの順に適用範囲が広くなるが，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に使用されるパターンの割合を図４に示す。なお、図では、最も基本的な文種別１と文種別３の文型パターン（いずれも述部数２）の場合についても示した。

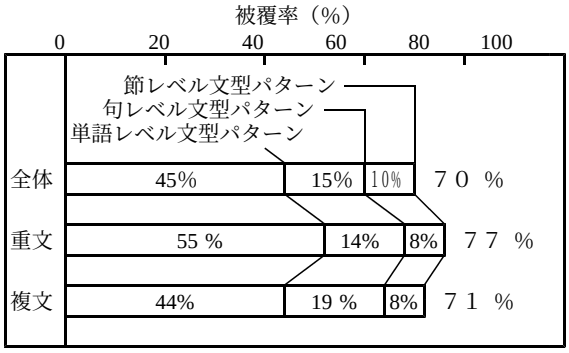

図４．文型パターン辞書全体での意味的な被覆率

図４から、文型パターン辞書を使用した機械翻訳では、重文・複文の約半分に単語レベルの文型パターン、残りの半分に句レベルと節レベルの文型パターンが使用され、最後の残りは、従来の翻訳方式が適用されることになると予想される。

７．まとめ

言語表現の非線形性に着目した言語表現モデルとそれに基づく文型パターン化の方法を提案し、日本語の基本的な重文、複文の表現に対して、単語レベル（12.3万），句レベル（8.0万件)、節レベル（1.2万件）からなる文型パターン辞書（合計21.5万件）を開発した．

パターン辞書作成の手順は以下の通りである。まず、基本的な日本語表現を含むと考えられる約30種類のドキュメントから、対訳例文100万件を取り出して対訳コーパスを作成し、２つまたは３つの述部を持つ重文、複文15万件を抽出した。次に、得られた対訳例文の線形要素を段階的に汎化して文型パターン辞書を作成した。

その結果によれば、汎化することができた線形要素は，自立語の場合62%（47万語／76万語）であっ

たのに対して、名詞句，動詞句などの句では、22%（10万件／46万件）、節は、きわめて少なく4%（1.1万件／27万件）であった．また、線形要素のみの文型パターンはXXXX件にすぎなかった。

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．この結果から，ほとんどすべての複文、重文は、複数の単文に分けて翻訳し後で結合しても、対訳例文に示されるような翻訳はできないことが分かった。

また、クロスバリデーションによる文型パターン辞書の評価実験では、単語レベル、句レベル、節レベルの文型パターンの意味的な被覆率は、それぞれ、43.3 %、55.2 %、50.1 %で、それらを組み合わせて使用するときは、92%であった。入力文に適合するパターンには、意味的に不適切なものも多い。それらを除いたときの意味的な被覆率は、70%であった。また、最も基本といえる述部数２の重文，複文の意味的な被覆率は、それぞれ、77%，71%であった．

文型パターンは，非線形な言語表現を対象としている。線形な表現の場合は従来の要素合成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．また、文型パターンは、意味を掬くい取るための網の目である。機械翻訳だけでなく、広く意味解析への利用が期待される。

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

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

ALT-JAWS: Morphological Analysser for Japanese,2002. http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws.html

有田潤：「ドイツ語講座II」南江堂, pp. 48-56, 1987.

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

Brown, R.D: Adding Linguistic Knowledge to a Lexical Example-Based Translation System, TMI 99, pp.22-32, 1999.

C. Fillmore: The mechanics of Construction Grammar. Berkeley Linguistics Society, 14. 35-55, 1988.

C. Fillmore, P. Kay, and M. Catherine O＇Connor: Regularity and idiomaticity in grammatical constructions: the case of let alone. Language, 64/3. pp. 501-39, 1988.

C. Fillmore, P. Kay, L. Michaelis and I. Sag: Construction Grammar, Stanford Univ Center for the Study, 2005.

H. A. Guvenir and Cicekli: Leaning Tanslation Rules from Examples. Information systems Vol. 23. No. 6, pp. 2325-363, 1988

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

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.

池原悟：「機械翻訳」、in「言語情報処理」pp. 95-148, 岩波書店，1998

S. Ikehara: Meanig Comprehension Using Semantic Patterns in a Large Scale Knowledge-Base, Proceedings of the PACLING'01, pp. 26-35, 2001

池原悟：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.

池原悟： 言語で表現される概念と翻訳の原理、言語と思考研究会、TL2003-25, pp. 7-12, 2003

Jung H.; Yuh S.; Kim T.; Park S.: A Pattern-Based Approach Using Compound Unit Recognition and Its Hybridization with Rule-Based Translation, Computational Intelligence, Vol. 15, No. 2, pp. 114-127, 1999

Kaji, H., Y. Kida & Y. Morimoto: Learning translation templates from bilingual text', COLING-92, pp. 672-678, 1992

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, Vol. 2, pp. 267-268, 2001.

計算機用日本語基本動詞辞書IPAL、情報処理振興事業協会 技術センター、1987

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

R. W. Langacker:FOUNDATION OF COGNITIVE GRAMMAR, Stanford University Press, 1987.

三浦つとむ：「認識と言語の理論」第１部～第３部，勁草書房，1967

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.

長尾眞：自然言語処理，岩波書店，1996.

E. Riloff: Automatically Generating Extraction Patterns from Untagged Text, AAAI-96, 1996

佐藤理史：アナロジーによる機械翻訳, 共立出版， 1997.

S. Sato: An example based translation and system, COLING-91, pp. 1259-1263, 1992.

S. Shirai, S. Ikehara, A. Yokoo and H. Inoue: The quantity of valency pattern pairs required for Japanese to English MT and their compilation. NLPRS '95, Vol. 1, pp. 443-448, 1995

E. Sumita, H. Iida, Experiments and prospects of Example-Based Machine Translation, 29th ACL, pp. 185-192, 1991

K. Takeda: Pattern-based Machine Translation, the 16th COLING, Vol. 2, pp. 1155-1158, 1996.

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

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.

S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao and A. Waibel: The CMU statistical machine translation system. Proceedings of MT Summit IX, pp. 402-409. 2003.

H. Watanabe and K. Takeda: A Pattern-based machine translation system extended by example based processing, 17th COLING, pp.1369-1373, 1998.

Watanabe, W. Kurohashi, S. and E. Aramaki: Finding structural correspondences from bilingual parsed corpus for corpus-based translation, COLING2000, 2000.

T. Watanabe and E. Sumita: Bidirectional Decoding for Statistical Machine Translation, Proceedings of COLING-02, pp. 1075-1085, 2002

R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen: Unsupervised Discovery of Scenario-Level Patterns for Information Extraction. In COLING-2000, 2000.