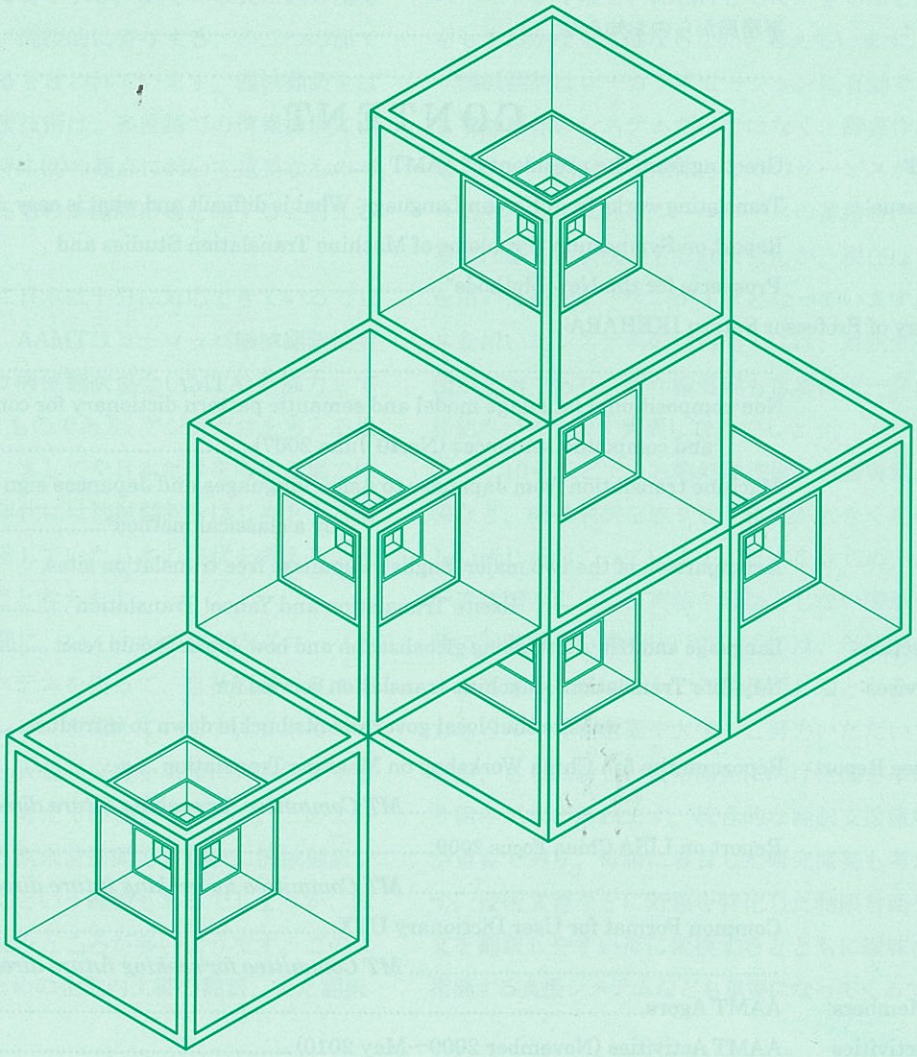
ISSN 1883-1818

AAMT Journal

Asia-Pacific Association for Machine Translation



June 2010 No.47

アジア太平洋機械翻訳協会

池原悟先生を偲んで

昨年末に鳥取大学の池原先生がお亡くなりになりました。まだまだご活躍・ご指導いただけると思っておりましたのに、残念でなりません。ここに池原先生の最後のご寄稿を再掲し、池原先生のご功績を偲びたいと思います。

池原先生との思い出をお話しするのには、私は適任ではないとは思いますが、数年前にヨーロッパの会議で池原先生と同じホテルに宿泊しており、先生と奥様が街を散策に出られるところにエレベータホールでお会いし、お元気になられたのだなと後ろ姿をお見送りしたことが思い出されます。最後にお会いしたのは鳥取大学で開催された昨年の言語処理学会の折で、長尾先生からの「お身体に気をつけて」とのお言葉をお伝えしたところ、長尾先生に気にかけていただけるとは光栄ですね、と喜んでおられました。

日本語語彙大系をはじめ、池原先生の数々の業績を私たちは十分には活かせていないように思います。機械翻訳の技術向上のためにも、先生のご研究を学び、発展させていきたいものです。

アジア太平洋機械翻訳協会会長　井佐原 均

追悼池原悟先生

非線形言語モデルと重文複文の意味類型パターン化

鳥取大学工学部知能情報工学科　池原　悟

1．研究の背景とあらまし

機械翻訳では、最近、統計翻訳方式が注目されているが、日英言語のように異なる言語族間で品質の良い翻訳を実現するのは難しい。実用システムの多くは、依然として要素合成法を基本とするトランスファー方式が中心で、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、パターン翻訳と用例翻訳は、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。このうち、パターン翻訳では、あらかじめ、パターン間の意味的な排他性を考えた大規模なパターン辞書を開発する必要のあるが、用例翻訳は、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、用例中の要素置換の可否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、結局のところパターン翻訳に帰着する。

また、最近、認知言語学のCognitive Grammar（ラネカー等）やConstruction Grammarの方法（フィルモア等）が注目されるが、構造的な意味の単位を決定する基準が曖昧で、工学的適用が困難である。

ところで、表現構造と意味を一体化した方法としては、すでに、日英機械翻訳システムALT-J/E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」が開発されている。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見られるが、この辞書は、動詞と名詞を中心とした表現（客体的表現）の意味を結合価パターンの形式にまとめたもので、助詞や助動詞の表現（主体的表現）の持つ非線形な意味が失われることが問題であった。また、従来の機械翻訳では、重文や複文のような長い表現の翻訳が特に不得手であるのに対して、これらの文に対する表現意味辞書は存在せず、依然として要素合成法に頼らざるを得ない点が問題であった。

そこで、筆者等は、これらの問題を解決するため、言語表現の構造と意味の関係に着目した「非線形言語モデル」を提案し、重文と複文に対する「意味類型パターン辞書」を開発した。この研究は、科学技術研究機構（JST）が推進する創造的基礎研究推進事業（CREST）の一つで、5年計画で実施したものである。

この研究では、日英対訳文（100万件）から、重文複文の対訳文（15万件）を抽出し、それに含まれる線形な表現要素を単語レベル、句レベル、節レベルの3段階で汎化することにより、合計22.7万件の対訳パターン辞書を実現した。得られたパターンは、一つ一つが、日本語表現の意味を掬い取る網のようなものである。そこで、重文複文の表す意味の体系を構築し、各パターンに該当する意味コードを付与することによって、パターンとそれによって掬い取られる意味の関係が明らかになるようにした。

以下、本稿では、この研究の概要について紹介する。まださまざまな問題を残しているが、既に80%近い被覆率が達成されており、実用的価値が期待される。

2．言語表現の持つ意味とは何か？

本研究では、構成部分の意味からだけでは全体の意味が説明できないような表現を「非線形表現」という。非線形表現は、それを構成要素に分解する過程で元の表現の意味が失われる。既に述べたように、従来の言語処理における最大の問題は、実際の言語では、非線形表現が多いことである。

このような非線形表現は言語に限らずさまざまな表現に存在する。心理学では、このような現象はゲシュタルトと呼ばれ、古くから主要な研究対象とされてきたが、言語処理では、非線形表現はやっかいなものとして避け、依然として要素合成法を前提とした研究開発が行われている。

ところで、分解する過程で意味が失われると言うことは、「表現の構造が意味を持つ」と言うことであるから、非線形表現の処理に挑戦するには、「表現構造の持つ意味とは何か」について明らかにする必要がある。そこで、まず、

この問題についての本研究の考え方を説明する。

（1）言語の本質は何か？

まず、「言語の本質は何か？」について考える。この問題はあまりにも初歩的で、今更と思われるかもしれないが、現実には、言語と記号の区別さえ明確に意識されない研究が見られる。

さて、言語は表現の一種であるが、絵画や音楽も表現の一種である。そこで、表現を図１に示すように、「感性的な表現」と「超感性的な表現」に分類する。前者は五感に訴える表現であり、後者は理性に訴える表現である。視覚に訴える絵画や聴覚に訴える音楽はいずれも前者の表現であるのに対して、記号（地図や道路標識等）や言語（自然言語）は、後者の表現である。

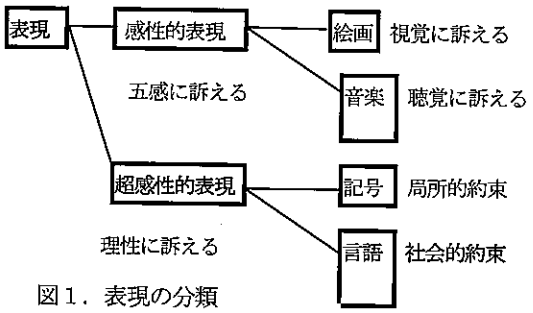


図１．表現の分類

ここで、問題となるのは、記号と言語との違いである。いずれも、表現と意味の関係についての約束を背景とする表現であるが、記号の場合、約束の適用範囲は局所的であり、原則として使用するに先立って定義を必要とする。これに対して言語の約束は、当該言語集団の中において自然発生的に成立した慣習であり、原則として定義することなしに使用できる。すなわち、記号と言語の本質的な違いは、社会的に共有された約束に支えられた表現であるか否かにあると考えられる。

ところで、言語が社会規範としての言語規範に支えられた理性的表現である以上、言語処理研究の第１の課題は、言語規範の特徴と性質を解明し体系化することだと言える。そこで、言語規範を「意味的な約束」（表現と意味の関係に関する約束）と「文法的約束」（語の並びに関する約束）に分けると、後者は前者の約束に共通した構造上の特長を取り出したものであり、二次的な約束と言える。

従来の言語処理は、二次的な約束を一次的な約束とみなし、本来の一次的な約束の体系化を棚上げしてきた嫌いがある。

（2）言語表現の意味とは何か？

一次的な約束を体系化するには、言語表現と意味の関係を明確にしておく必要があるが、言語表現の意味については、言語哲学的にも混乱した状況にあり、定説が存在しない。

このような現状の中で、三浦つとむは言語過程説において「関係意味論」を提唱している。この意味論は、「言語表現とそれに対応づけられた話者認識の関係」を表現の意味とするものである。「対象」、「話者の認識」、「表現」、「聞き手の解釈」など言語実体のいずれも意味とはせず、それらの関係を意味としている点でユニークな説と言えるが、「実際に使用された表現に対してのみ定義されること」、「意味は、それぞれの表現に固有で、客観的であり、解釈によって変化するものでないこと」が、計算機処理にとって好都合である。そこで、以下では、「関係意味論」に基づき意味的約束の性質を明らかにする。

（3）概念とはどのようなものか？

関係意味論に基づき、言語表現の意味は、言語表現とそれに対応づけられた話者の認識との関係であるとする。表現に対応づけられる認識は概念だと考えられるが、そのように断定して良いかについては、解決しておくべき問題がある。そこで、概念と表現の関係について検討する。

さて、概念は一般に、「対象の持つ特殊性を普遍性の側面から採り上げた認識の単位」である。言い換えると「対象の持つ個別的特徴を捨象し、必要不可欠の要素を統一体として捉えたもの」と言うことができる。

言語表現と概念の関係を考えるには、「認識としての概念」と「言語規範としての概念」の違いを明確にする必要がある。前者は、話者の認識の中で形成された概念であり、後者は、各言語において規範として成立している概念である。言語表現では、話者認識の中で概念化に成功した内容が必ずしもそのまま言語で表現できるわけではない。言語規範として成立している概念を介してのみ初めて言語で表現することができる。この違いは、三浦つとむが「概念の二重性」と呼んでいるものである。

言語意味処理の研究において、どちらの処理を目指すかによって大きな違いが生じる。前者の処理では、「世界知識に関する知識ベース」が必要となるのに対して、後者は、「表現と概念に関する知識ベース」があれば実現可能と見られる。筆者はこの違いに着目して、前者の処理を「意味理解」、後者の処理を「意味解析」と呼んで区別しているが、本研究では、後者の概念を意味処理の対象とする。

（4）言語表現は概念を表すか?

さて、「果たして、言語表現は概念化された認識を表すと言って良いか」の問題に戻ることにする。言語表現において単語が概念を表す点については異論のないところであるが、問題は、「複数の単語からなる表現も概念を表すか」について、従来、明確な議論が見あたらないことである。

時枝誠記は、言語過程説において本居宣長の「玉の緒の理論」を引き継ぎ、日本語表現を「客体的表現」と「主体的表現」に分類したが、前者を「概念化された客体の表現」、後者を「概念化されない主体の感情や意思の表現」と説明している。これは、言語において概念化されない認識の表現を認めたものである。三浦つとむは、これを修正し、どちらの表現も概念化された認識の表現だとしている。

言語が理性的な表現であることを考えると、概念化に成功しない認識が言語で表現できるとする考えには矛盾がある。筆者の考えによれば、「客体的表現」と「主体的表現」との違いは、対象と主体との間に対峙関係が存在するか否かの違いによって説明できる。そこで、本研究では、「客体的表現」、「主体的表現」共に概念化された認識の表現であると考え、概念化されない認識は言語では表現されないと考える。

以上の議論により、複数単語から構成される言語表現も概念を表すと考え、本研究では、言語表現から見た概念を以下の２種類に分類する。

■単一概念＝単一の単語で表現できるような概念
■複合概念＝複数単語の表現（句、節、分など）によって表される概念

この分類は、言語表現の形式に着目した分類でありあくまで相対的である。「単一概念」は、内部構造を意識しないまでに抽象化された概念であるのに対して、「複合概念」は、内部構造を意識した概念、すなわち適切な単語で表現できないため、その内部構造まで意識して複数単語の表現に対応付けざるを得ない概念である。

言語では、新語の意味を文中で定義して使用したり、一度述べた内容（複合概念）を代名詞（それ、これ、等）で表現したりするが、これらは、複合概念を単一化する仕組みの一部である。

以上から、認識、概念、表現の関係は図２のようにまとめることができる。

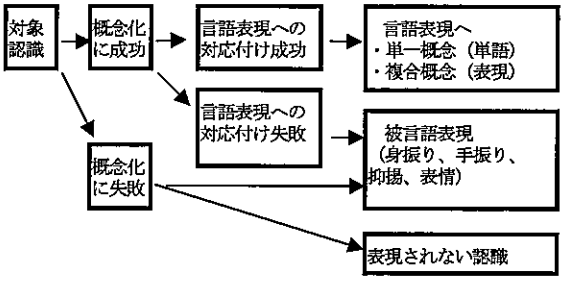


図２．認識・概念・表現の関係

（5）言語表現の構造も概念を表すか？

前節では、複数単語からなる表現が表す概念を「複合概念」と呼ぶことを述べた。そこで、いよいよ、「表現構造も概念を表すと言えるか？」について考える。

既に述べたように概念は、「対象の特殊性を普遍性の側面から採り上げた認識」である。対象の持つ普遍性をどのレベルで捉えるかによって、概念間に階層関係が生じる。例えば、「太平洋は日本海より広い。」、「富士山は大山より高い。」に共通する概念を「相対比較」と呼び、「相対比較」と「絶対比較」に共通する概念を「比較」と呼ぶことができる。

これに対して、言語表現の構造は、表現の形式を表すものであり、個別的な言語表現を抽象化したものである。個別的な概念を表す表現を抽象化すれば、より上位の概念を表す表現になると考えられる。このことから、それぞれのレベルの概念にはそれに対する表現の形式が存在し、逆に、表現の形式には、それによって表される概念が対応すると考えることができる。

以上から、個々の言語表現と同様、表現構造も概念を表すと言って良いことが分かる。但し、以上の議論は、概念化に成功した認識のみが、言語で表現されることを前提としている。概念には、表現や表現構造が対応するが、逆に、言語表現や表現構造のどのような部分を切り出してもそれに対応する概念が存在する訳ではないので注意が必要である。

３．非線形言語モデルとパターン化

本章では、このモデルの概要と重要な性質のいくつかについて説明する。

なお、前章で述べたように、表現の表わす内容は概念である。従来の言語処理では、表現の表わす内容を意味と呼んでいる場合が多いので、それに倣って、以下では、表現の表す概念を意味と呼ぶが、それは本来の意味でない点に注意する。

3．1　言語表現の線形性と非線形性

「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)。

ところで、話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「表現構造の意味」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を「非線形な表現」と定義する。

但し、定義１の「表現構造の意味」は、前章の結論から、「抽象化された複合概念」である。上記の定義を現実の言語表現に適用するには、これを記述する方法を明確化する必要がある。

ところで、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。

図３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。単一概念である「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。

3．2　非線形要素の基本的特徴

前節の定義では、線形要素は以下の５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、

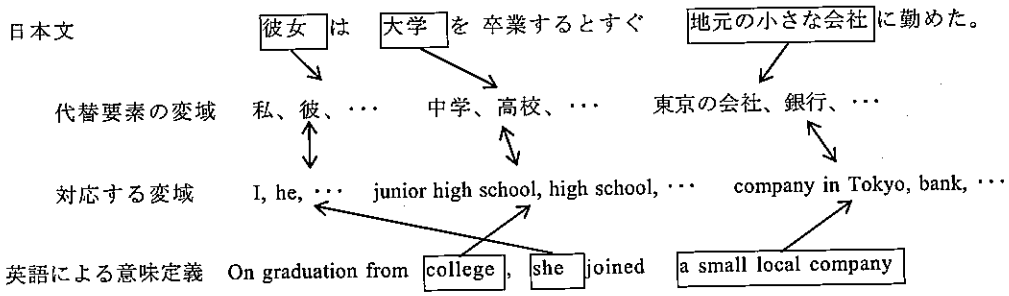


図３．線形要素の例

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＞線型要素の相互独立性

線型要素は代替要素に置き換えても全体の意味が変わらないことが条件である。この条件は各線型要素毎に成立しなければならないから、一つの表現構造の中に複数の線型要素がある場合は、それらの線型要素間の関係は相互独立でなければならない。

＜特徴３＞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４＞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５＞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。

３．３　非線形言語モデル

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４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５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４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、非線形言語モデルでは、言語表現は、非線形要素と非線形要素から構成される。

３．４　なぜ非線形な表現をパターン化するか

図２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する表現意味辞書を構築すれば、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、非線形表現を記述する方法であるが、

(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対象にパターン化する。

４．文型パターン辞書の開発

４．１　パターンの記述方法

本章ではパターン記述方法の概要を説明する。詳

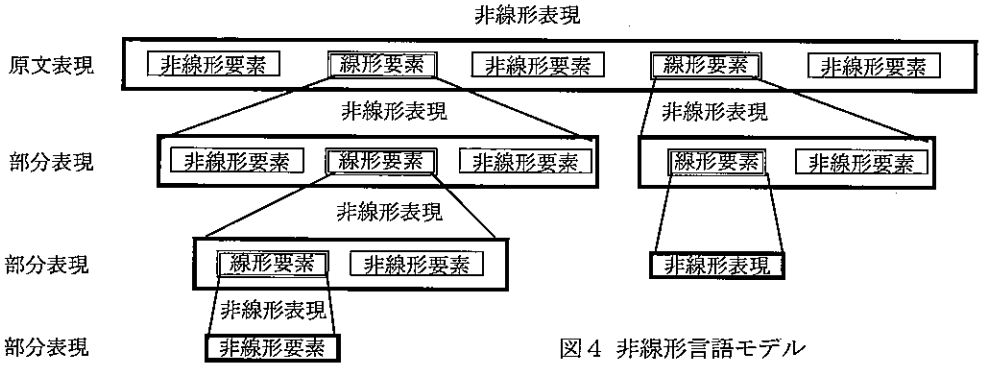


図４　非線形言語モデル

細は、「意味類型パターン記述言語仕様書」をご覧頂きたい。

（１）パターン化の原則

前章では、日本語表現の意味を英語表現を用いて定義することを述べた。この典型的な例は、日英対訳文である。そこで、本研究では、日英対訳例文の汎化によってパターン対を作成することとし、パターン記述言語を開発した。

ところで、汎化の対象となるのは、線形な表現要素である。前章の定義から、日英対訳例文において、線形要素には下記のものがある。

（１）英語側に対応する要素があるもの

（２）英語側に対応する要素はないが、削除しても英語表現は変化しないもの

（３）元の対訳表現にはないが、それを挿入しても、倦関係が変化しないもの

（４）語順など構造的な変更を行っても英語側の構造が変化しないもの

そこで、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こととした。

（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
1	字面	日本語文字、英語文字
2	変数 (17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 (157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 (10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、表現には現在形でしか使用できないもの（非線形）や現在形、加工形の双方で使用できるものなど（線形）がある。このような時制、相、様相などの表現の線形性と非線形性は、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。

４．２　文型パターンの作成方法

（１）手順の概要

まず、日英対訳例文100万件を収集し対訳コーパスを作成した。次に、その中に含まれる述部２つまたは３つ重文、複文を取り出し、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。対象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。

なお、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。

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。

（２）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。

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(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 fignificantly toxic so that great caution must be taken with its use.

＜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。

4．3　作成した文型パターン

（1）生成された文型パターン数

文型パターンの記述例を表２に示す。また、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

表３．文型パターンの種類と異なりパターン数

	単語レベル	句レベル	節レベル	合　計
重文	61,171	39,243	18,173	118,587
複文	48,123	32,049	5,778	85,950
重複文	12,610	8,146	1,524	2,280
合計	121,904	79,438	25,475	226,817

に示す。単語レベルでは、元の標本文とほぼ同じだけのパターンが作成されたが、句レベル、節レベルになるにつれ汎化が困難となり、得られたパターンは急速に減少した。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、73.9%、句では24.0%であったのに対して、節ではきわめて少なく14.83%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、多くの場合、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。

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)	734,528	542,833	73.9 %
句	463,636	111,359	24.0 %
節	267,601	39,714	14.8 %

4．4　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

②実験はクロスヴァリデーションの方法とする。

このうち②は、パターン作成に使用した例文を入力文に使用するが、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方法である。

文型パターンの被覆率は以下の2つのパラメータによって評価した。

- 統語的被覆率（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
- 意味的被覆率（C）：入力文に対して、意味的に正しいパターンが1つ以上存在する割合

（2）統語的被覆率の飽和特性

文型パターンの数と適合率の関係を図５に示す。曲線は、下から順に、単語レベル、句レベル、節レベルのパターンの飽和特性を示す。

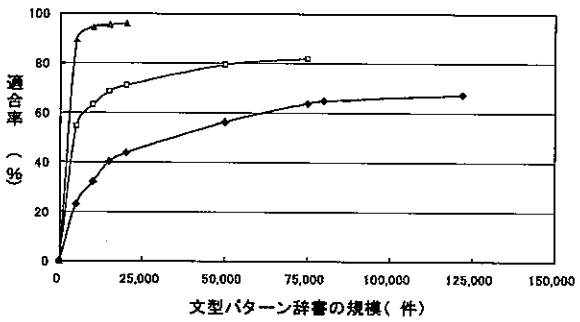


図５．文型パターン適合率の飽和特性

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。過去の研究によれば、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。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。

（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。この表がら、文型パターン辞書全体では、構文上、入力文のほぼ全体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78%である。

表５．文型パターン辞書の被覆率特性

パターン種別	平均適合パターン数	統語的被覆率(R)	意味的被覆率(C)
単語レベル	63.2 件	62.6 %	36.7 %
句レベル	520.3 件	82.7 %	55.0 %
節レベル	1241.7 件	96.4 %	67.0 %
合　計	1825.2 件	97.8 %	78.0 %

なお、述部数2の重文、複文の意味的な被覆率は、それぞれ、82.5%、71%であった。

５．あとがき

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（12.2万）、句レベル（7.9万件）、節レベル（2.5万件）からなる文型パターン辞書（合計22.7万件）を開発した。

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73.9%であったのに対して、名詞句、動詞句などの句では、24.0%、節はきわめて少なく14.8%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、ほとんどすべての複文、重文は、複数の単文などの要素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品質の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、36.7 %、55.0 %、67.0 %で、それらを組み合わせて使用するときは、78%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、82.5%、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。

なお、本稿では、述べなかったが、本研究では、「意味的等価変換方式」（本ジャーナル、No.33、pp.1-7、2002.3を参照)の実現に向けて、重文複文の意味分類体系を構築し、すべてのパターンに意味分類コードを付与した。その詳細については、別の機会にご紹介する予定である。

付記1

本研究の成果の研究利用は原則無料です。詳細は、「鳥バンク」（http://unicorn.ike.tottori-u.ac.jp/toribank/）にてご案内する予定です。多くの方々のご利用を期待しています。

付記2

本研究が契機となって平成19年度文部科学大臣表彰科学技術賞（研究部門）を受賞しました。今まで多くの方々に支えられ、意味解析の研究を続けてくることができたお陰です。今後の機械翻訳技術の発展を願うと共に、その活動を支えてこられた多くの方々に深くお礼を申し上げます。