非線形な表現構造に着目した重文と複文の日英文型パターン化

池原　悟[†]　　阿部 さつき[††]　　徳久 雅人[†]　　村上 仁一[†]

あらまし　要素合成法を基本とした従来の機械翻訳方式の限界を突破する方法として，非線形な言語表現の構造を意味のまとまる単位にパターン化した文型パターン翻訳方式が期待される．本論文では，重文と複文を対象に，この方式の実現に必要な文型パターン辞書を試作した．具体的には，100 万件の日英対訳コーパスから２つ又は３つの述部を持つ重文と複文合計 15 万件を抽出し，単語レベル（12.8 万件），句レベル（10.5 万件），節レベル（1.3 万件）の３種類のグループからなる文型パターン辞書（合計 24.6 万件，異なり 22.1 万件収録）を作成した．各文型パターンは，いずれも形態素解析によって得られる文法情報を用いて記述することとし，対訳標本文に含まれる線形な表現要素を半自動的に変数化，関数化することなどにより作成したものである．従来，大規模な文型パターン辞書の開発は，文型パターン間の意味的排他性実現の困難性と膨大な開発コストが問題となるため，適用対象を限定するなど小規模な実現例しか見られなかった．しかし，今回の試作によって，ほぼすべての標本文（99%）が多くの線形要素（平均４～５カ所）を持つことが分かった．また，それらの要素を半自動的に関数化，変数化を行うことにより，文型パターンの開発コストは人手に頼る方法の約 1/10 に削減できた．これにより，実験的検討に必要な規模の文型パターン辞書を構築することができた．

キーワード:　機械翻訳，文型パターン，言語知識ベース，要素合成法，非線形要素，汎化

Japanese to English Sentence Pattern Generations for Semantically Non-Linear Complex Sentences

Satoru Ikehara[†] and Satsuki Abe[††] and Masato Tokuhisa[†] and Jin'ichi Murakami[†]

Abstract : In order to breakthrough the limitation of the conventional method based on Compositional Semantics, it is expected to realize a new translation method based on Sentence Patterns in which non-linear structures of linguistic expressions are represented as semantic units. This paper proposes the way to judge the linearity or non-linearity of linguistic expressions based on their definitions and how to generate sentence patterns from huge bilingual corpora. According to this method, three kinds of sentence patterns such as "word level", "phrase level" and "clause level" are generated in this order from Japanese to English corpus. In the experiments, 150,000 sentence pairs for complex and compound sentences are extracted from one million sentence pair corpora, and 128,000 patterns, 105,000 patterns and 13,000 patterns for each of three revels were generated from these sentence pairs. Due to the clarifications of decision process, the generation processes of the sentence patterns were mostly automated by using the results of morphological analysis and these 246,000 sentence patterns have been obtained in a year.

[†] 鳥取大学工学部　Faculty of Engineering, Tottori University
[††] NTT アドバンステクノロジ株式会社　NTT Advanced Technology Corp.,

自然言語処理　Vol. 2　No. 3　July　1995

KeyWords: *Machine Translation, Sentence Pattern, Linguistic Knowledgbase, Compositional Semantics, Nonlinear Expression, Generalization*

1　はじめに

従来，研究開発されてきた機械翻訳システムは，ほぼいずれもトランスファー方式を基本としている．この方式は,「原文の構文構造を目的言語の構造に変換する過程」と「原文の各要素を翻訳する過程」を持ち，訳文は両者の結果を合成することによって得られる点に特徴がある（長尾 1996；長尾，黒橋，佐藤，池原，中尾 1998）．これは，構文構造と表現の意味を別々に変換するものであり，表現の構造と意味の関係が線形であることを前提とした要素合成方式が基本となっている．しかし，現実の言語表現には非線形なものが多く，表現が構成要素に分解される過程で全体の意味が次第に失われ，目的言語を生成する過程で復元できなくなることが問題であった（池原 2001）.

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」を実現することが重要である．文構造とその意味を一体的に扱う仕組みとしては，古くから「文型パターン翻訳」の方法が試みられてきた．文型パターン翻訳は「テンプレート翻訳」とも呼ばれている．パターンに適合する入力文に対して品質の良い訳文が得られることから，多くの商用システムでトランスファー方式と併用する形で実現されてきた．最近では「翻訳メモリ」とも併用される傾向にある.

しかし，これらの文型パターン翻訳で使用されている文型パターン数はいずれも少なく（200〜300 パターン程度），特定の狭い分野の文書に適用される例が多い．これは，パターン作成のコストが大きいこと，また，パターン数を増やすとパターン間の意味的な相互作用が増加して翻訳精度が低下することによるためと考えられる.

これに対して，既に，構造と意味の関係を考慮した「多段翻訳方式」が提案されている（池原，宮崎，白井，林 1987）．この方式は，原言語表現の構造を意味を失わないように目的言語に対応づける仕組みとして「結合価パターン」を使用している．パターンの意味的排他性の問題は，精密な意味属性体系を使用することで解決しており，単文レベルの翻訳において精度の良い訳文が得られている（金出地，徳久，村上，池原 2003）．しかし，複文（埋め込み文を持つ文），重文（接続のある文）の持つ非線形性が扱えないこと，また，原文に対して単一の目的言語表現が対応づけられる仕組みであるため，文脈に応じた表現選択ができないことが問題として残されている.

これらの２つの問題を解決するため，最近，言語表現の意味類型化を基本とする「意味的等価変換方式」が提案された（池原，佐良木，宮崎，池田，新田，白井，柴田 2002）．この方式は非線形な言語表現の構造を意味的に類型化（衛藤，池原，池田，佐良木，新田，柴田，宮崎，白井 2003）した「意味類型パターン辞書」を使用するが，この辞書を構築するためには，やはり大規模な「文型パターン対辞書」を作成する必要がある.

文型パターンは，言語表現の非線形な構造を取り出してパターン化したものであるが，大規模な対訳コーパスからこのような文型パターンを作成するには，与えられた言語表現のどの要素が線形要素であり，どの要素がそうでないかを判断する基準を明確にし，作業手順化することが重要である．

ところで，線形性と非線形性は，表現構造と意味の関係に対して定義されるものであるため，現実の言語表現に適用するには，与えられた各表現の意味の定義を必要とする．すでに，言語表現の意味については，関係意味論の立場から概念の二重性に着目した検討（池原 2003）が行われており，また，言語表現構造の線形性，非線形性の問題については，工学的立場からの検討（池原 2004）が行われてる．

本検討では，その結果に基づいて日本語表現の意味を英語表現で表すこととし，大量の日英対訳例文から非線形な表現構造を取り出して文型パターンを作成するための作業基準と作業手順を定めた．また，それに従い，重文・複文を対象に構文レベルで 24.6 万件の日英文型パターンを作成したので，その結果について報告する．

以下，第２章では，機械翻訳の立場から，各文要素が文全体に対して線形であるか否かを判断するための原則を示す．第３章では，汎化作業の方針と単語レベル，句レベル，節レベルの３段階の文型パターン作成のための作業項目とその基準を述べる．最後に第４章では，重文と複文の対訳標本文から作成された文型パターンの内容を示す．

2　文型パターン化の原則と方針

本章では，与えられた言語表現の線形性，非線形性を判定するための原則とそれに基づき文型パターン化を行うための方針を示す．

2.1　言語表現における線形要素と非線形要素

文献（池原 2003）によれば，言語表現の線形性は以下の通り定義される．

【定義１：言語表現の線形要素】

特定の概念 (「複合概念」) を表現するための表現構造の要素のうち，他の要素（意味の異なる要素でも良い）に置き換えても表現構造全体の意味（「複合概念」[1]）が変わらないとき，その要素をその表現構造の「線形要素」と言う．

【定義２：表現構造の線形性と非線形性】

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」と言い，１つ以上の非線形要素を有する表現構造を「非線形な表現構造」という．

[1] (池原 2003) によれば，言語表現は概念化された話者の認識を表し，単語は「単一概念」を表すのに対して複数単語からなる表現は「複合概念」で表す．すなわち，概念化されていない認識は言語 (「直接的表現」) では表現されないから，以下では「言語表現の意味」と「言語表現の表す概念」を同義で使用する．

本稿ではこの定義に従って対訳例文の線形要素を汎化し，文型パターンを作成することとする．この定義を現実の言語表現に適用する方法と注意すべき点を以下に示す．

(1) 英語による表現の意味の定義

通常，単語が単一概念を表現するのに対して，上記の定義は,「句，節，文等の表現は話者の認識の中で形成された複合概念の表現（池原 03）である」ことを前提としている．

そこで，各言語は複合概念を表すための様々な形式を持っていることに着目し，与えられた日本語表現の意味（複合概念）を英語表現によって記述することとする[2]と，定義１の「表現構造全体の意味が変わらない」ことは,「対応する英語表現の構造が変わらないこと」と読み替えることができる．

その結果，与えられた日英対訳用例において，着目する日本文の文要素が線形であるか否かを判定するには，それを他の文要素に置き換えたとき，対応する英文全体の表現構造が変化するか否かを調べればよいことになる．

(2) 線形要素の重要な性質

文型パターン化を考える上で，上記の定義の持つ意味は以下の通りである．

＜線形要素の制約条件＞

第１は文型パターンの線形要素の置き換え範囲（値域）の問題である．定義１は，他の要素に置き換えても表現構造全体の意味が変化しないような要素を線形要素としているが，これは実際にどんな要素に置き換えても良いことを意味しない[3]．元々日本語側で見て，意味をなさない表現になるような置き換えはできず，線形要素と言えども，置き換え可能な範囲には一定の制約がある．

＜要素の選び方と全体の線形性＞

第２は表現要素の線形性と表現全体の線形性の関係である．定義２によれば，すべての要素が線形な場合に限り表現は線形だとされている．これは，表現全体の線形性は，その要素分解の方法に依存して決まることを意味している．また，要素の単位を指定すれば，線形，非線型の区別は一意に決定できることから，汎化の程度に応じた文型パターンが作成できる．

＜文全体の非線形性と文要素自身の非線形性＞

第３は，線形，非線形の区別は表現の部分と全体の関係を言うものであり，線形要素だと言ってもその要素自身が線形であることを意味しないことである．線形要素の内部構造は非線形であっても良い．このように，線形，非線形の分類が再帰的な構造を持つことは，長文の構造が複数の非線形構造の組み合わせで解析できることを意味する点で大変重

---

[2] 言語表現の意味をいかなる記法で記述しても計算機から見れば，単なる記号に過ぎないから，意味記述言語は，表現能力があり，相互矛盾のない体系であればよい．その点，自然言語は表現能力の高い言語である．そこで，目的言語を使用して原言語の意味記述を行うこととするが，この方法は，機械翻訳システム構成上，便利な方法と考えられる．

[3] 例えば,「私は彼より背が高い.」の文において,「私」を「あなた」に置き換えても，この文が「２者比較」と言う「複合概念」を表す点での意味は変化しない．しかし,「私」を「川」や「月」などに置き換えると，表す複合概念が変化する以前に文としても成り立たなくなる．

要な性質である．

2.2　文型パターン化の原則

前節の定義に従って対訳コーパス中の例文に含まれる線形要素を抽出し，それを汎化することにより，日英文型パターン対を作成する．以下，日英対訳文から日英対訳文型パターンを作成するための原則について述べる．

(1)　文型パターン化の対象としない対訳文

現実に得られる対訳例文の品質は様々である．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．

但し，文型パターン化の目的が非線形な言語表現の意味を正しく翻訳することにあることから，意訳された対訳例を文型パターン化することは極めて重要である．そのような例文では，無理な汎化はせず対訳原文をそのまま文型パターンとしても良い．

(2)　文法レベルでの文型パターンの記述

本稿では文型パターン記述用の言語として，文献（池原，宮崎，佐良木，池田，白井，村上，徳久 2003；池原，村本，徳久，村上，宮崎，佐良木 2004）で提案された「文型パターン記述言語」を使用するが，文法レベルの情報を使用して文型パターンを記述することとし，変数の変域に対する意味的な制約条件は付与しない．また，語順の変更や文型要素の移動可能指定の機能も使用しない[4]．

これは，現段階では意味的制約条件付与の必要性と必要な意味の粒度などが不明なためである．意味レベルでの文型パターン記述の必要性とその方法などについては，文法レベルで記述された文型パターンの被覆率特性が明らかになった段階で検討する．

(3)　必須要素と任意要素の区分

対訳例文の要素を以下で示すような「必須要素」と「任意要素」に分類する．

＜必須要素＞　：日本語文型パターン内にそれがないと対応する英語文型パターンが定義できない要素を言う．

＜任意要素＞　：日本語文型パターン内にそれがなくても英語文型は決定できる要素で，文型パターン定義に使用するか否かによってさらに以下の２つに分類する．

- 「原文任意要素」：削除されても対応する英訳文は変化しない要素で，文型パターンでは陽に示されない．
- 「パターン任意要素」：陽に示しておかないと訳語や訳語挿入位置の決定が困難な要素で，文型パターンの一部として陽に示される．

[4] 提案されている「文型パターン記述言語」では，変数の意味的制約条件の記述方法や文型要素の出現順序の可変性の指定方法等も規定されているが，本検討では，第１段階として文法的な属性のみの情報で記述された文型パターンを試作することを課題としている．

なお,「必須要素」と「任意要素」は，いずれも字面でも良いし後で述べるような変数や関数を含む表現でも良い.

(4) 変数化とその範囲

変数化する対象は，単語（自立語，複合語を含む），句，節の３種類の表現でいずれも線形な文要素である．これに応じて英語表現中の対応する文要素も変数化する.

今回の文型パターン試作の狙いは，(i) 総合的に被覆率の高い文型パターンが得られるか，(ii) 文型パターン相互間の意味的独立性が確保できるかの２点についての指針を得ることである．相反するこれら２つの目標を調和的に実現するため，対訳例文を段階的に汎化することによって文型パターンを作成することとし，変数化された文要素の単位に応じて，文型パターンを「単語レベル」,「句レベル」,「節レベル」の３種類のグループに分類する．以下，変数化の基本原則を示す.

まず，変数化の対象となる文要素であるが，変数化する文要素はいずれも別途翻訳して英語文型パターンに埋め込めるもので，それ自身が線形である必要はない．変数化判定の原則は以下の通りである.

(i) 英語側に対応する訳語（句と節を含む）を持つ場合

日本語と英語で変数化される要素は必ずしも文法的に同じ属性である必要はなく，品詞や活用形が異なっていても良い．日英の対応する部分の品詞が異なる場合は，品詞変換の関数を使用して変数化する.

(ii) 英語側に対応する訳語を持たぬ場合

英語側に対応する訳語を持たない要素は，前項の分類によって「原文任意要素」か「パターン任意要素」かのいずれかと判定する.

(iii) 日本語と対応しない英語文型パターン要素

逆に，英文中に日本文に対応づけられないような要素については，日英対訳文の意味的な関係を調べ，文脈なしにその対訳関係が成り立つなら，英語パターンの要素として残し，そうでない場合は削除する.

次に，変数化する要素の範囲であるが，入力文と日本語文型パターンとの照合時の結果，変数にバインド（代入）され，英語文型パターンに持ち運ばれる範囲である．問題となるのは用言性の文要素であるが，汎用性の高い文型パターンとするため，用言の変数化においては,「語幹＋活用形」の範囲を変数化し，時制，相，様相の情報を表す自動詞などについては，別途定められた関数を使用して記述する．従って，通常，文型パターンでは用言の活用形は指定されないが，指定の必要な場合は，活用形指定関数を使用する.

また，日英文型パターン間での変数の対応関係を明確にするため，まず，日本語文型パターン内で使用する変数には通し番号を付与する．これに伴い，英語文型パターン内の変数には，日本語文型パターン内の意味的に対応する変数と同じ番号を付与する．日本語文型パターンで使用された変数のすべてが英語文型パターンで使用されている必要はない．なお，同一の文型要素の変数化では，同一の変数番号を使用するものとする.

(5) 関数化などによる汎化

付属語（辞と辞相当語）要素のうち線形なものは，各種の関数および選択記号によって汎化するが，入力文と文型パターンとの照合の段階で意味的な曖昧性の発生しない方法（主として字面指定の関数または選択記号）で記述する．

2.3 文型パターン化の個別的方針

文型パターン化における個別的な方針を示す．

(1) 作業自動化の可能性の追求

膨大な対訳例文から文型パターンを能率良く生成するため，対訳例文の形態素解析と構文解析情報を使用し[5]，文型パターン作成の半自動化を目指す．そのため，機械的な変数化と関数化ができるよう，文型パターン作成に先立って元となる対訳標本文は形態素解析し，解析誤りは人手で修正しておくこととする[6]．

(2) 文型パターン照合の容易性

従来の構文解析では解消困難な構文多義の問題を解決することも文型パターン翻訳方式の目標の一つである．そこで，入力文と文型パターンの照合では構文解析は使用せず，入力文の形態素解析結果を使用することとし，文型パターンは入力文の形態素解析結果との照合において曖昧さの発生しない方法で記述する．これに伴い，形態素解析で文法的，意味的に解釈の確定しない表現要素については文法的，意味的な指定は行わず，字面もしくは字面の指定される関数を用いて文型パターンを記述するものとする．

(3) 変数書き換えによる文型パターンの汎化

句レベルの文型パターン化では，対訳標本文中の線形な句（既に単語変数化された表現を含んでも良い）を変数化するが，より汎化の範囲を拡大するため，単語レベルで変数化した単語変数についても句変数に置き換え可能なものを探して句変数化する．例えば，名詞を表す単語変数 N を名詞句変数 NP に置き換えることができれば，より適用範囲の広い文型パターンが生成できたことになる．

(4) 離散記号による文型パターンの汎化

原文任意要素の定義から分かるように，機械翻訳で使用される文型パターンは，パターン要素のすべてが入力文に含まれていることが条件となるが，逆に入力文には，文型パターンに定義されない要素を含むことも認められる．

そこで，文型パターンがより多くの範囲の入力文に適合するようにするため，入力文と文型パターンの照合を制御するための離散記号を使用し[7]，可能な限りこの記号を使用

---

[5] 文型パターン照合では入力文の構文解析情報は使用しないが，文型パターンの作成では使用する．

[6] 構文解析の精度は形態素解析に比べて十分とは言えないため，予め標本文すべての解析結果を人手で修正するには多大なコストが必要となる．そこで，構文解析プログラムは，句変数化，節変数化などで構文情報が必要となる標本文に限って限定的に使用する．

[7] この記号（スラッシュ記号 ”／ ”）は，文型パターン要素間（但し，文節境界）に挿入するもので，この記号がある位置は，原文任意要素が挿入されても良いことを意味する．

した文型パターン化を行う．

(5) 字面レベルでの表記の揺れの吸収

日本語は表記の揺れが多く，このことが，文型パターンとの照合で大きな問題となる．この問題を解決するため，入力文と文型パターンとの照合の段階で形態素解析プログラムの持つ標準表記認定機能を利用することを前提に，原則として標準表記を用いた文型パターン記述を行うこととするが，同時に，格助詞，助動詞，副詞（いずれも相当語を含む）等について，可能な限り標準表記と異表記の関係をまとめた表を準備し，字面グループを指定する関数や選択記号を使用して使用可能な表記を指定する．

3　文型パターン化の具体的方法

前章で述べた原則と方針に従って，表１で示すような，単語レベル，句レベル，節レベルの文型パターンを順に作成する．また，各レベルにおける作業項目一覧を表２に示す．以下，これらの作業の内容について述べる．

表 1: 各レベルにおける文型パターン化の内容

	汎化の内容
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まで行ったもの． （１）名詞，動詞，形容詞，副詞などの自立語の変数化 （２）線形で文型上不要な要素を任意化と文型の骨組みとなる要素の抽出 （３）字面要素についてのグループ化
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の拡大 （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, 従属節の変数化による重文複文の基本構造のパターン化 （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

3.1　単語レベルの文型パターン化の方法

すべての対訳例文に対して，単語レベルの文型パターンを１パターンづつ作成する．文型パターン化の作業は,「原文任意要素の削除」,「自立語の変数化」,「述部語尾表現の関数化」,「パターン任意要素の指定」,「表現要素のグループ化」,「各種加工」の６種類から構成される．以下，それぞれの具体的内容について述べる．

表 2: 文型パターン化作業項目とその内容

#	汎化規則の分類			汎化規則の内容	備考
1	単語レベルの汎化	原文任意要素の削除		原文任意要素を削除する	
2		自立語の変数化	名詞一般	・格要素の名詞の変数化, ・述語名詞の変数化, ・数詞の変数化（数詞として識別する）, ・複合名詞の単一名詞化	名詞機能語は対象外．底の抽象名詞 (とき，原因，理由，等) も対象外.
3			複合名詞	述語単独動詞の変数化	機能語として使用された和語動詞は対象外
4			動詞一般	単独動詞連体形, と連用形の変数化	
5			複合動詞	複合動詞の変数化	主動詞のみ変数化
6			形容詞・形容動詞	述語形容詞・形容動詞の変数化	
7				単独形容詞形容動詞連体形, 連用形の変数化	
8			副詞	文修飾，用言修飾の副詞の変数化	変数化の可否に要注意
9			品詞変換関数の適用	品詞変換関数を使用する. （表記法の例）$V(N2)$，$N(V3)$,	活用形表記法も併用
10		述部語尾表現の関数化		英語アスペクト情報の関数化	（例）$.ing$, $.ed$, $.er$, $.st$ などの関数を使用
11				英語様相情報の関数化	
12		パターン任意要素の指定	名詞修飾語	単独の動詞連体形，形容詞, 形容動詞連体形，連体詞の任意化	英語文型の決定に不要な線形要素をバッカス記号 #n[ ] で囲む．任意化する部分は，変数を含んで良い.
13			用言修飾語	単独動詞連用形，単独形容詞, 単独形容動詞連用形の任意化	
14				副詞的用法の名詞（昨日，今日）の任意化	
15			その他	英語に訳出されるその他線形要素の任意化	
16				英語に訳出されない要素の任意化	日本語側を [ ] で囲む
17		表現要素のグループ化		同種の格助詞相当語や副詞をグループ化	$[A \mid B]$ の記法による
18		各種加工	主語の補完	英語主語に相当する名詞が日本語側にない時，日本語パターンに補う．例） $[N1]$ 英語側のパターンは $[N1 \mid \text{he}]$ の形式で記述.	パターン適用時は，$[N1]$ 要素はなくても良い
19			冠詞の削除	英語パターン内の変数された名詞の冠詞を削除	定まった冠詞は対象外
20	句レベルの汎化	句の変数化		名詞句，動詞句，形容詞句，形容動詞句, 副詞句の変数化	単純な構造を持つものを対象とする
21		適用範囲拡大	現在形変換	現在形でも使用される文型パターンを 対象に past() を削除する.	
22			丁寧表現の標準化	丁寧表現をフラットな表現に変更する. ＜注＞英語訳出されない丁寧表現が対象	接頭辞「お」,「御」等を削除
23			N を NP へ拡張	名詞変数 N を名詞句変数 NP に置換. それに伴い不要となったパターン 任意要素を削除.	可否判断要注意
24			V を VP へ拡張	名詞変数 V を名詞句変数 VP に置換.	可否判断に要注意
25			機能語の拡張	格助詞を同種の格助詞相当語に置き換え	
26				述部の語尾を同種語尾パターンに置き換え	
27				同じ意味の副詞等をパターンに追加	$(A \mid B)$ の記法による
28	節レベルの汎化	節の変数化		名詞節，連体節，連用節，引用節, 並列節の単一変数化	
29		節レベル関数の使用		日本語節を英語句に変換する関数の使用	
30				英語節構造を指定する関数の使用	

3.1.1 原文任意要素の削除

以下の２の条件を満たす要素を「原文任意要素」と判定し，原文から削除する．

(i) 対応する文要素が英文にあり，日英双方の文からそれらを取り去っても対訳関係が成り立つこと．

(ii) 着目する文要素を単独で英訳することができ，英訳文に組み込めること．訳語選択，訳語位置決定の困難な要素は，対象外．

また，日本文と対応関係を持たない英文中の要素も原文から削除する．なお，(i) の条件は満たすが，(ii) の条件を満たさないものが,「パターン任意要素」である．

＜例＞ 下記の例 (1) では、対訳関係にある「その後」と”after that”は、英語文型構成上重要でないので、原文任意要素として削除。例 (2) の「一日中」は、”all day ”と対訳関係にあるが、英語文型上の訳出位置情報が重要と見られるため、削除せず、文型パターンの一部とする。

(1) その後、彼女は靴を脱いで祭壇に上がっていった。
After that, she took off her shoes and stepped up to the alter.

(2) 一日中 歩き回ったので足が棒になった。
After walking all day, my legs feel like logs.

3.1.2 自立語の変数化

語彙的な用法で使用された線形な自立語を以下の順序で変数化する．なお，機能的用法で使用された自立語は変数化しない．

(i) まず，日本語表現内の自立語に対応する訳語が英訳文中の同一の品詞の語（複合的な表現でも良い）として存在し，対訳辞書などによってその対応関係が決定できるものを日英同一の変数（変数名と変数番号共に同一）に置き換える．

(ii) 残された日本語表現内の自立語のうち，線形なものを変数化する．この場合，英語側に意味的に対応する自立語がなくても良い．また，英語側の自立語で，日本語側に対応する自立語を持たないものは変数化しない．

＜例＞ 下例のように、予め対訳辞書を使用し、単語の意味的な対応関係を決めておき、それが線形要素である場合はそのまま変数化する．

手段は目的を達成するためのものだ。$\longleftrightarrow$ A means is for attaining a purpose.

$N1$　$N2$　$V3$	$N1$　$V3$　$N2$
$N1$ は/$N2$ を $V3$ ためのものだ。	$N1$ be for $V3.ing$ $N2$.

以下，自立語変数化の個別的な方法を示す．

(1) 名詞の変数化

語彙的用法の名詞と複合名詞を名詞変数 N に置き換える．但し，数詞，時詞は NUM，$TIME$ を使用する．また，機能語として使用された名詞「の，もの，こと，人，とき，場合，原因，理由」などは変数化しない．

複合名詞の場合，英語文型に影響を与えるような接頭辞，接尾辞を複合名詞一部として名詞変数 N に含めるような変数化は行わない．そのような場合は，複合名詞を構成する要素の一部を字面で残し,「大手金融会社→大手 N」などのように部分的に変数化しても良い．なお，NN のような変数化はパターン照合を難しくする割に効果がないので行わない．

(2) 動詞の変数化

語彙的用法の動詞の「語幹＋活用語尾」を動詞変数 V に置き換える．機能動詞「ある，なる，いる，くる，する，行く，言う」などは，変数化しない．

$V + V$ の複合動詞は，アスペクトの意味を有することが多いので，全体を 1 変数化せず,「V 始める」,「V てみる」，などのように主動詞側を変数化し，アスペクト情報を担う動詞は変数化しない．また，$N + V$ のような複合動詞は，N，V 共に変数化しても良いし，いずれか，片方を字面としても良い．なお，命令形と連用形など文型パターン選択で活用語尾が手がかりとなる文型パターンについては，活用形指示関数を使用する．

(3) 形容詞，形容動詞の変数化

形容詞，形容動詞の変数化は動詞と同様である．従って，形容動詞は「名詞＋だ（助動詞）」とはしない．

(4) 副詞の変数化

名詞の場合と同じである．但し，変数化の基本原則 (i) を満たさない場合が多いと思われるので，注意が必要である．

(5) 品詞変換関数の適用

日本語側の文型パターン要素が，英語文型パターンで文法属性の異なる自立語に対応するものを対象に「変数関数」を使用した文型パターン化を行う．対象となる表現部分は日本語側の N，V がそれぞれ英語側で V，N になるもので，英語側の要素を変数関数 $V(N)$，$N(V)$ で表現する．

3.1.3 述部語尾表現の関数化

動詞，形容詞，形容動詞に接続する助動詞および助動詞と助詞の連鎖表現（連語）を「形式指定関数」を用いて書き換える．書き換える際の注意点は以下の通りである．

(i) この段階では，線形な時制，相，様相の情報の任意化は行わない．

(ii) 従って，日本語側，英語側の文型パターンにおいて，これらの情報は,「形式指定関数」を使用するが，それがないときは，関数を新設するか字面関数を使用する．

なお，対応する英語文型パターンの箇所で必ずしも日本語文型パターンで使用した関数を使用する必要はない．使用する関数の種類は英語側で独自に決定する．

＜例＞　買う。→ V。　買った。→　$V.kako$。　浴びている。→　$V.teiru.kako$。
忘れられない。→　$V.rareru.hitei$。
開けておいて下さい。→　$V.teoku.tekudasai$。
元気になったようだ。　→　$AJV.joutaihenka.kako.suitei$。

3.1.4　パターン任意要素の指定

日英対訳文型パターンを見比べ，下記の線形要素を「パターン任意要素」として指定する．

(i)　単独の動詞，形容詞，形容動詞の連体形と連体詞
(ii)　単独の動詞，形容詞，形容動詞の連用形
(iii)　英語に訳出されるその他の線形要素
(iv)　英語に訳出されない要素

但し，英語文形パターンでこの要素が指定されたときは，対応する日本語文型パターンに同一の番号を持つパターン任意要素が指定されていなければならない．

＜例＞　下例では、下線部分がパターン任意要素である。文型パターンでは、日英の対応関係が分かるよう番号（#n）をつけて表示される。
彼は 社会の 福祉に貢献した功を認められた。
→　$N1$ は／#2［社会の］$N3$ に／ $V4.kako$ ／功を認められた。
He won public recognition for having contributed to public welfare.
→　$N1$ won public recognition for $V4.pft$ to #2[public]$N3$.

3.1.5　表現要素のグループ化

助詞，助詞相当語，または，副詞などの字面のうち，同一の意味で異なる表記を持つものを対象に，選択記号を用いて置き換え可能な表現として指定する．指定される要素は必須要素とパターン任意要素を共に含んでよい．

＜例＞　下例は、下線部分の表記の揺らぎを認めるため、選択記号を使用して置き替え可能な表記を列挙したものである。
自分 ひとりで何でも やるのが彼の主義だ。
→　$N1$（ひとりで｜一人で）（何でも｜なんでも）やるのが／#2［$N3$］の／ $N4.da$
It is his principle to do anything whatever for himself.
→　It is #2[$N3.poss$]$N4$ to do anything whatever for $N1.reflex$.

3.1.6　各種加工

前節までで得られた日英文型パターンに対して「ゼロ代名詞の補完」,「冠詞の削除」を行う.

(1)　主語の補完

対訳例文において英語パターンで主語となる名詞が日本語側では記述されていない（ゼロ代名詞化されている）場合，日本語側パターンのゼロ代名詞の部分にパターン任意要素記号を使用して主語を補完する[8].

なお，この処理では，重文や複文の主節と従属節が同一の主語を持つか否かに注意する.

＜例＞　下例で、＜ $N1$ は＞の要素と（$N1$｜I）の要素が対応する。通常＜＞記号の部分の名詞変数 $N1$ に該当する名詞がバインドされ、英語側でその訳語が主語として使用されるが、＜＞の部分に該当する要素がなくても文型照合に成功する。但し、その場合、英語側の主語は、”I ”が使用される。

車を持つだけの資力がない。　→　＜ $N1$ は＞ $N2$ を／ $V3$ だけの／ $N4$ がない。

I don’t have means enough to have a car.　→　($N1$｜I)do not have $N4$ enough to $V3$ $N2$.

(2)　冠詞（a, the）の削除

英文生成過程では，名詞に対する冠詞の必要性とその種類および数を決定することが重要であるが，これらは英語訳語に依存して決まることが多いため文型パターンで厳密に定義することは困難である．そこで，英語パターンを使用した英文生成では，形態素調整処理において名詞変数の直前に冠詞のついていない名詞変数について冠詞と数の表現の必要性を判定するための処理を起動することとする.

従って，英語文型パターンに対して，文の意味上，定まった冠詞を要求する文型では，変数 N の前の冠詞を字面のまま残し，それ以外の場合は，冠詞はすべて削除する.

3.2　句レベルの文型パターン化の方法

句レベルの文型パターン化では,「句の変数化」,「現在形変換」,「丁寧表現の標準化」,「名詞から名詞句への拡張」,「動詞から動詞句への拡張」,「機能語の拡充」の６種類の汎化作業を行う.

＜例＞　名詞句変数と動詞句変数を使用した例を示す。

(1)　３塁の不安定な守備 が そのチームのアキレス腱 だ。 →　$NP1$ が $NP2$.da。

The uncertain fielding of the third baseman is the Achilles’s heel of that team. →　$NP1$ is $NP2$.

(2)　あの役人 は地位を悪用して 金を儲け た。 → $NP1$ は／ $N2$ を／ $V3$.$site$ $VP4$。

$NP1$　　$N2$　　$V3$　　$VP4$

[8] 例えば，日本語文型パターンで省略された主語を［$N1$ は］の形式で指定し，対応する英語側のパターンでは，該当する箇所を［$N1$｜he］の形式で変数化したとする．この場合，この文型パターンを使用した翻訳では，日本文に $N1$ に該当する要素のないときは，英語側のパターンでは ”he ”が使用される.

That government official earned money by abusing his position.
$NP1$ $VP4$ $V3$ $N2$
→ $NP1\ VP4.past$ by $V3.ing\ NP1.pron.poss\ N2$

（注）$NP1.pron.poss$ は $NP1$ の代名詞の所有格を意味する。

(1) 句の変数化

変数化する句は線形な「名詞句」,「動詞句」,「形容詞句」,「形容動詞句」,「副詞句」である．以下，それぞれ変数化する句の範囲を示す.

＜名詞句の変数化＞

格助詞と名詞の連鎖した表現を名詞句変数化の対象とする．また，動詞，形容詞などの単独用言の連体形により修飾された名詞の表現では，英語側に意味的に対応する部分があるものについて，修飾語と名詞の範囲を名詞句変数化の対象とする．それ以外の名詞句は変数化の対象としない．１つ以上の格要素を持つ用言によって修飾された名詞の表現は名詞句とはせず,「動詞句（形容詞句）＋名詞」と解釈する.

なお，名詞句の変数化では，以下に示すように形容詞の部分や一部の字面や単語変数として残し，残りの部分を変数化しても良い.

- 「茶色の大きな山」 →「N の NP」
- 「大変大きな山」 →「$ADV1\ AJ2\ NP$」，但し,「大変 NP」はだめ.
- 「前期試験の結果」 →「N の NP」，又は,「NP の N」，又は「NP」．但し,「前期 NP」はだめ.
- 「私の母の財布」 →「私の NP」, 又は,「NP の N」，又は,「NP」．但し「N の NP」はだめ.

＜動詞句の変数化＞

主格（が格，は格）を除く１つ以上の格要素を伴った用言について，動詞とその配下にある格要素すべての範囲の表現を動詞句変数化の対象とする．なお，主格を含む場合は，格要素と動詞を含む全体を節と解釈する.

＜その他の句の変数化＞

形容詞句，形容動詞句の変数化は，動詞句の場合に準じる．また，副詞句の場合は，名詞句の場合に準じる.

(2) 丁寧表現の標準化

英語文型パターンに影響のないことを確認して，日本語文型パターンの丁寧表現をフラットな表現に変更する．例えば,「〜する (V)，〜します (V ます)」,「〜しない ($V.not$)，〜しません」では，いずれも前者に縮退させる.

(3) 名詞から名詞句への拡張

修飾部を持たない名詞変数 N のうち，句に書き換えて良いものを選び NP に書き換える．また,「$AJ1\ N2$」や「$N1$ の $N2$」等のように連体修飾部を持つ名詞は，全体を NP に置き換える．後者の置き換えでは，パターン任意要素として任意化された助詞結合型名

詞句の主名詞を除く部分や連体詞，単独形容詞の連体形，単独動詞の連用形を含む名詞はすべて NP に置き換える．

(4) 動詞から動詞句への拡張

以下の条件に従って，動詞変数 V もしくは動詞句を動詞句変数 VP に置き換える．

(i) 英語側の主語に対応しない格要素を１つ以上を持つ動詞句であること．

(ii) 英語側の該当する表現に主語がないこと（英語側に主語のある表現では，既に単語レベルで日本語文型パターンの任意要素として，主語が補われている）．

(iii) 英語側も対応する句変数が使用されること（関数引数として使用されても良い）．

(iv) なお，日英の該当する動詞とそれを修飾するすべての格要素をまとめて一つの句変数に汎化すること．すなわち該当する格要素の一部を残してはいけない．

具体的な手順は以下の通りである．まず，修飾部を持たない動詞変数 V のうち，句に書き換えて良いものを選び，VP に書き換える．格要素や副詞に修飾された動詞の場合は，英語側の文型パターンとの対応関係を明示しなければならない要素部分を除き VP に置き換える．また，すでにパターン任意要素に指定されている副詞，単独形容詞の連用形，単独動詞の連用形は，修飾先の動詞と共に VP に書き換える．

なお，時制，相と様相に関する関数は削除しない．複合動詞も単語レベルでの汎化のままとする[9]．

(5) 機能語の拡充

ここまでに得られた文型パターン対の日本語文型パターン側を対象に，(i) 格助詞，格助詞相当語の部分と (ii) 述部の語尾表現で置き換え可能な単語もしくは連語を考えて，要素選択記号（α｜β｜…）の形式のパターンとする．追加する単語，連語は，パターン全体の意味を変えないもので，通常よく使われる範囲の語でよい．

3.3　節レベルの文型パターン化の方法

節変数 CL を用いた英語文型パターンの汎化を行う．簡単な例を示す。

＜例＞　畳の上に１時間も座っていたので足がしびれた。→　$VP1.teiru.kako$ので／$CL2.kako$。
　　$VP1$　　$CL2$

After sitting on the tatami floor for an hour, my feet were asleep.
　　$VP1$　　$CL2$

→　After $VP1.ing$, $CL2.past$.

(1) 変数化する節

以下の条件を満たす線形な名詞節，連体節，連用節，引用節，並列節を変数化の対象とする．

[9] 例えば，２つの従属節「彼は学校をさぼったため」と「彼は学校をさぼるため」では，前者は原因節，後者は「目的節」の意味になるから，句レベルの汎化で標準型を現在形とすることはできない．前者のパターンは過去形（$.ta$ 関数を使用），後者のパターンは現在形を指定する必要がある．

(i) 用言とそれに付随する格要素の範囲の表現（「パターン任意要素」が含まれても良い）のうち，英語表現の主語に対応する格要素を持つこと．

(ii) 英語側に対応する節形式の要素が存在し，その意味は日本語側の節の意味に対応すること．

(iii) 意味的対応関係は，文脈等の情報に依存しないこと．また，節内に英語文型全体を左右する要素を持たないこと．

(2) 変数化する節の範囲

変数化する範囲は,「命題命名のレベル」[10]の単文（核文とも言う）とする．すなわち，節変数に代入される値は，用言（体言述部の場合は体言）とそれに付随する格要素，副詞要素の範囲とし，時制，相，様相の情報を示す関数類はそのまま残す.

3.4 文型パターン化作業の半自動化

半自動化を考慮した作業手順を図１に示す．図１の手順で重要な点は以下の通りである．

(1) 原文の形態素解析情報の準備

対訳文の形態素解析結果から，文型パターンを半自動的作成できるようにするため，対訳文に対する形態素解析結果の誤りを人手修正し，正しい形態素解析結果を準備する．

(2) 日英文の自立語の意味的対応づけ

上記の結果と日英対訳辞書を使用して，日英例文中で意味的に対応関係を持つ自立語を自動的に対応づけ，それを変数化候補とする．なお，意味対応関係を持つ部分の文法属性（品詞など）が異なるものについても変数化の候補とする．

(3) 関数化対象表現の抽出

関数化の対象となる字面を予め関数毎にグループ化し，グループ化された対応表と (1) の結果を参照して，関数化候補箇所を自動抽出する．

(4) その他の各種要素の記述支援

任意要素，補完要素の指定，冠詞と丁寧表現の汎化などについても可能な限り判断規則を設け，その候補となる要素を自動抽出する．

これらの方法で抽出された汎化候補に対して，線形要素であることが機械的に明確に判定できるものについては，機械的に変数化，関数化を行う．また，機械的な判断規則が得られないものは候補のまま残し，言語アナリストの判断にゆだねる．言語アナリストが判断に迷うものについては，設計者との協議により個別に扱いを決定する．

[10] 日本文は下記の４つのレベルの表現の入れ子構造で捉えられる.（１）対象領域（命題）のレベル：(i) 事態命名のレベル（時制の関与なし），(ii) 個別的現象のレベル（相，様相が関与），（２）主体領域（様相レベル：(iii) 判断のレベル（断定の助動詞，形容詞「ない」），(iv) 表現・伝達のレベル（各種助動詞）．このうち，節変数化は，(i) のレベルを対象とする．

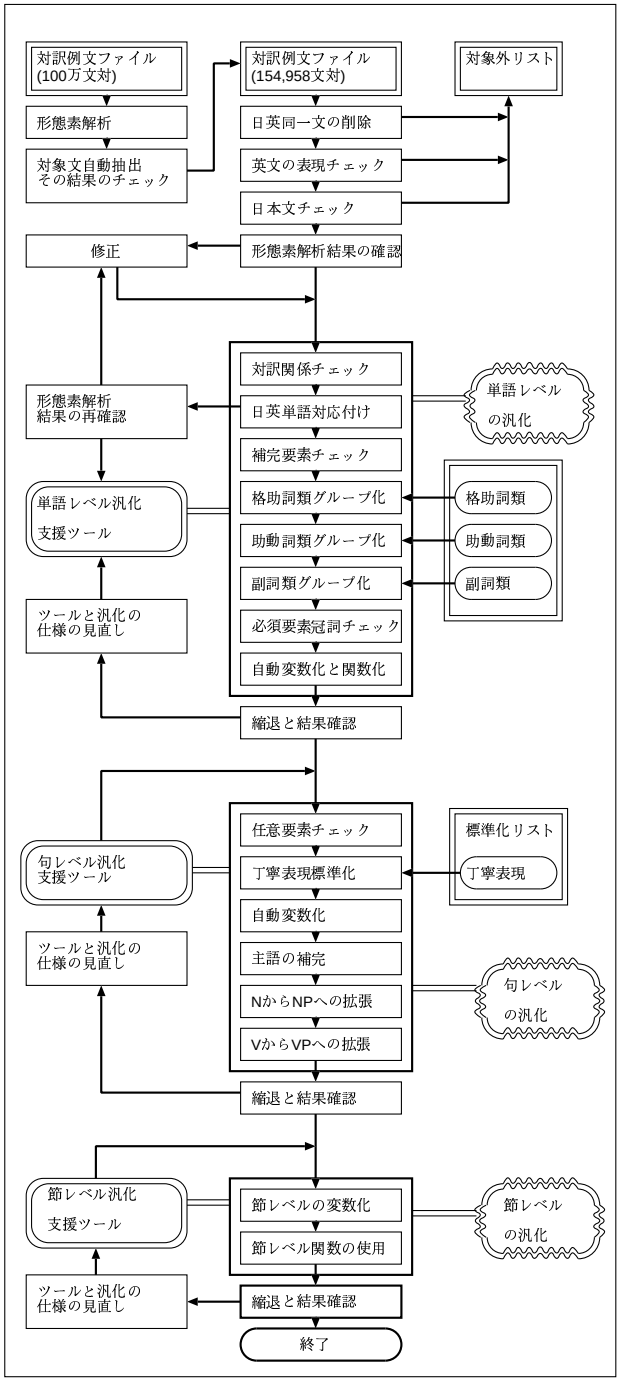

図 1: 文型パターン化の作業手順

4　重文と複文の文型パターン化

述部を２つ又は３つ含む重文と複文（記述文）を対象に前章で述べた方法により文型パターン化を行った．重文，複文を文型パターン化の対象としたのは以下の理由による．すなわち，日本語表現のうち，単文の非線形構造については，既に日本語語彙大系（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）においてまとめられており，高品質の翻訳が可能となっているのに対して，重文複文の非線形構造については，類似の知識ベースがなく，訳文品質は依然として低いレベルにとどまっているからである．また，述部の数を２と３に制限したのは，重文，複文と言えども，現実の文では，４個以上の述部を持つ文全体が非線形であることは少なく，そのような文は，述部２または３の文型パターンに分解して翻訳できる可能性が高いと考えられるためである．

4.1　対象例文と作業の状況

(1)　対象とする日英対訳例文

まず，辞書や日本語教材をはじめとする約 30 種類の対訳データファイルから作成した 100 万文の対訳コーパスから，重文，複文 15.5 万件の対訳文を機械的に抽出した．その中には，会話文や文脈依存の訳文が含まれていたため，それらを人手で振るい分け，12.9 万件を標本文として３種類の文型パターンを作成した．表３に抽出した対訳文と文型パターン化の対象とした標本文の内訳を示す．

対象文に含まれる単語の種類と数を表４に示す．また，対訳標本文の平均文字数などを以下に示す．

日本語原文　：平均文字数／文＝ 23.3 字（最大 148 字）
　　平均形態素数＝ 12.9 個／文（最大 63 個）

英文訳文　：平均単語数＝ 10.3 語／文（最大 59 語）

表 3: 対訳標本文数と作成した文型パターン数

文種別	説明	抽出した対訳文数	対象とした標本文数
文種別１	文接続１カ所を持つ文	72,018	57,235 (44%)
文種別２	文接続２カ所を持つ文	7,292	6,196 (5%)
文種別３	埋込み文１つを持つ文	54,931	46,907 (36%)
文種別４	埋込み文２つを持つ文	6,688	5,986 (5%)
文種別５	文接続と埋込文各１つを持つ文	14,029	12,389 (10%)
－－	合計	154,958	128,713 (100%)

(2)　文型パターンの適切性の検証

文型パターン化作業の結果を確認するため，文型パターン照合プログラムを用意し，作成された全文型パターンに対する照合実験を行った．作成した文型パターンの総数と不適切文型パターン数の推移を表５に示す．

表 4: 品詞毎の出現回数

#	種別	形態素数	
		延度数	異り語数
1	名詞	417,886	56,861
2	本動詞	223,178	10,324
3	補助動詞	51,918	271
4	形容詞	31,681	915
5	形容動詞	19,587	2,562
6	副詞	39,051	3,191
7	連体詞	32,585	731
8	接続詞	3,146	77
9	感嘆詞	147	60
10	接頭辞	1,068	110
11	接尾辞	1,749	336
12	助動詞	165,251	236
13	助詞	465,811	349
14	記号	121,555	32
−	合 計	1,574,613	76,055

表 5: 不適切文型パターン数の推移

種別	作成した	不適合パターン数	
	文型パターン数	初期段階	現段階
単語レベル	128,071	58,194 （45.4%）	165 （0.13%）
句レベル	104,619	18,643 （17.8%）	1,810 (1.7%)
節レベル	13,031	4,638 (35.5%)	2,381 (18.3%)
合　計	245,721	81,475	4,356 (1.8%)

この実験では，文型パターンの作成に使用した標本文を入力文として文型パターン辞書を検索し，適合する文型パターンの中に，自分自身から作成された文型パターン（「自己パターン」と言う）が含まれるかどうかを確認した．

その結果，最初の段階では，単語レベル，句レベル，節レベルで，それぞれ 45%，18%，35%の文型パターンに誤りがあることが判明したが，誤りの多くは，自動変数化プログラムに組み込まれた規則と文型パターン記述仕様との不整合に起因するもので，人手修正の必要な文型パターンの記述誤りは約 5,000 件であった．節レベルでは，まだ 18%の不適合パターンを残しているが，この大半は，文型照合プログラムの機能不足により照合に失敗しているもので，文型パターン記述の誤りは少ないと思われる．

以上の結果，文法レベルにおいて約 24.6 万件の文型パターンをほぼ 1 年間（3.4 人年）で作成することができた．人手作業に頼る場合[11]に比べて，作業工数は，約 1/10 に減少したものと推定される．

[11] 日英対訳標本文から単語レベル，句レベル，節レベルの文型パターンの組を作成するのに 30 分かかると推定して，(30 分/標本文）× 15 万標本文＝ 450 万分＝ 37.5 人年と見積される．なお，いずれの場合も作業者としては，英語の素養のあるベテランの日本語アナリストの動員が必要である．

4.2 作成された文型パターンの例と数

作成した文型パターンの例を表6に示す．また，その数の内訳を表7に示す．

表 6: 文型パターンによる非線形構造の記述例

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
単語レベル	文型パターン	それは/$N1$ に/あるまじき /$N2$.da。	Such $N2$ be unseemly for $N1$.
	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.
	文型パターン	#1[$N1$ は]/$N2$ の/$V3$/こと/#2[ある/程度は] /$V5$.	#1［$N1$ | I］ can $V5$ what $N2$ $V3$ #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/$V1$.$temiru$.ta が/$N2$ が/$NP3$.da。	All things $V1$.$past$, $N2$'s $NP3$.
	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［$N1$ は］/$NP2$ も/$V3$.nai とは/$VP4$.$gimu$/ ことだ。	It is $AJP(VP4)$ that #1[$N1$ | you] should $V3$.not $NP2$.
	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	［$N1$ は］/$CL2$ とは/$V3$.nai.$kako$。	#1［$N1$ | I］did $V3$.not $CL2$.$past$.
	適用例	彼があれ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.
	文型パターン	$CL1$.$teiru$.nai.da と/$N2$ は/$VP3$.$kako$。	$N2$ $VP3$.$past$ that $CL1$.not.
	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.

表 7: 対訳標本文数と作成した文型パターン数

文種別	標本文数	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	57,235 (44%)	56,883 (3,305)	45,241 (7,885)	6,343 (822)	108,467 (12,012)
文種別2	6,196 (5%)	6,179 (99)	5,079 (127)	424 (7)	11,682 (233)
文種別3	46,907 (36%)	46,684 (2,676)	38,654 (7,722)	3,807 (622)	89,145 (11,020)
文種別4	5,986 (5%)	5,973 (84)	5,307 (223)	874 (63)	12,154 (370)
文種別5	12,389 (10%)	12,352 (178)	10,338 (313)	1,583 (32)	24,273 (523)
――	128,713 (100%)	128,071 (6,342)	104,619 (16,270)	13,031 (1,546)	245,721 (24,158)

単語レベルのパターン化では，標本文数 128,713 文中，642 文は変数化される部分がないため，字面のみの文型パターンとして残され，128,071 件の文型パターンが得られた．このうち 6,342 件 (5%) は，互いに同一のパターンとなったため，異なり文型パターン数は 121,729 件である．

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，104,619 件の文型パターンが得られた．このうち，16,270 件 (16%) は，同一の文型パターンとなったため，異なり文型パターン数は，88,349 件である．

これに対して，節レベルで作成された文型パターンは 13,031 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は非線形要素であり，汎化困難であること，すなわち，重

文，複文構造の９割程度は，節の翻訳結果を組み合わせる従来の方法（要素合成法）では良い翻訳ができないことを示している．

なお，全文型パターンのうち，同一のものは 1,546 件 (11.9%) であった．記述レベルの違いによる文型パターンの縮退の程度を見ると，単語レベル（5%）に比べて，句レベル（16%）は３倍以上大きい．これはほぼ予想されたところで，汎化するにつれて文型パターンは縮退することを意味している．

以上，３段階の文型パターン化で得られた文型パターンの総数は，字面パターンを含め，日本語文型パターン 24.6 万件であった．そのうち同一の文型パターンで縮退されるものは 24,157 件（9.8%) であるので，異なり文型パターンの合計は，221,564 件である．

4.3　変数・関数の使用頻度

(1) 変数の使用頻度

表８に文型パターン全体で使用された変数の種類と頻度を示す．

表 8: 変数の使用回数

変数種別	使用回数 (割合)		
	単語レベル	句レベル	節レベル
N	303,319 (64.2%)	138,033 (43.4%)	10,135 (30.8%)
$TIME$	8,527 (1.8%)	5,187 (1.6%)	529 (1.6%)
NUM	6,036 (1.3%)	2,314 (0.7%)	189 (0.6%)
V	101,484 (21.5%)	48,036 (15.1%)	4,254 (12.9%)
REN	21,241 (4.5%)	2,158 (0.7%)	127 (0.4%)
ADV	11,491 (2.4%)	7,631 (2.4%)	603 (1.8%)
AJ	10,950 (2.3%)	6,193 (2.0%)	425 (1.3%)
AJV	9,473 (2.0%)	6,273 (2.0%)	434 (1.3%)
VP	—	58,908 (18.5%)	2,838 (8.6%)
NP	—	40,629 (12.8%)	1,985 (6.0%)
AJP	—	1,341 (0.4%)	78 (0.2%)
$AJVP$	—	935 (0.3%)	37 (0.1%)
$ADVP$	—	117 (0.0%)	8 (0.0%)
CL	—	—	11,280 (34.3%)
合計	472,521 (100%)	317,755 (100%)	32,922 (100%)

表４から，変数化の対象となり得る自立語は，名詞，本動詞，形容詞，形容動詞，副詞，連体詞を合わせて約 76 万語であったのに対して，表８によれば単語レベルで変数化されたものは 47 万語であるから，自立語の 62%が変数化されたことになる．文型パターン当たりで見ると自立語 5.9 語中の 3.7 語が変数化され，非線形要素として字面のままとなったものは 2.2 語である．

句レベルでは１つの句変数の中に複数の自立語が縮退された場合を含め，59 万語 (78%) が変数化されている．また，標本文に含まれる節は 28.2 万件であったのに対して，それが節変数化されたものは，1.13 万件（４%）にすぎない．

このことから，自立語，句，節のそれぞれ 38%，22%，96%が非線形要素であったこと

になる．これを標本文数と対比すると，１文あたり，単語レベルでは平均２語あまり，句レベルでは平均 1.3 語程度の自立語が変数化されていない．このことから，標本文には，節レベルではもちろん，自立語でも非線形なものがかなり多く存在していることになる．

(2) 関数の使用頻度

文型パターン全体で使用された関数の種類と頻度を表９に示す．この表から，文型パターン当たりの関数使用回数は単語レベルでは平均 0.7 回，句レベルでは 0.95 回，節レベルでは 1.5 回である．汎化のレベルが上がるにつれて関数の使用頻度が増えている点から，使用される変数の種類の違いだけでなく関数化の点でも汎化が進んでいることが分かる．

(3) 線形な文型パターンと非線形な文型パターン

既に述べたように初期の段階で汎化不能として文型パターン化の対象外としたものが 624 件あったが，実際に作成した文型パターンのうち関数や変数が使用されず字面のままとなったものは単語レベルで 302 件存在した．いずれも線形要素を持たないと判定されたものである．対象標本 12.9 万件中，これらの割合が 0.72 %であることから，殆どの標本文（99%以上）は一つ以上の線形要素を持っていたこととなる．

また逆に，字面を含まない文型パターンは，単語レベルで 15 件，句レベルで 401 件，節レベルで 155 件であった．このことから，いずれの場合も線形な文型は少ないが，特に，単語レベルで少ないことが分かる．

以上のように，殆どすべての重文・複文の構造は要素合成法の適さない非線形であるとは言え，その殆どすべてが１つ以上の線形要素（平均４～５カ所）を持つ．これは，多様な言語表現がカバーできるような文型パターンが開発できる可能性を示しており，文型パターン翻訳は，用例翻訳の限界を超える方法として期待される．

4.4　汎化による文型パターンの同一化

(1) 同一化する文型パターンの割合

文型パターン数と被覆率の関係を考えると，標本量が増大するにつれて，同一化するパターンの増大することが期待される．しかし，実際に 12.9 万件の標本文で得られた文型パターンで同一化したものは，単語レベル，句レベルでそれぞれ約 5%，約 16%でかなり少ない．

これは，機械翻訳において，入力文のすべての要素が解釈できるような文型パターンを網羅的に準備することは困難であることを意味するが，機械翻訳で使用される文型パターンは必ずしも入力文の全要素を含むものである必要性はない．すなわち，機械翻訳文で使用可能な文型パターンは，その要素のすべてを含む入力文でなければならないが，逆に，入力文から見れば，すべての要素が文型パターンの要素に一致している必要はない．

従って，開発すべき文型パターン辞書の必要規模を推定するには，部分的な適合条件を

表 9: 関数の使用回数

関数名 (40 種)	使用回数 (割合)		
	単語レベル	句レベル	節レベル
ta	33,155 (38.42%)	33,138 (37.57%)	5,614 (36.23%)
teiru	9,737 (11.28%)	11,090 (12.57%)	2,126 (13.72%)
reru	8,663 (10.03%)	8,750 (9.92%)	1,282 (8.27%)
da	7,879 (9.13%)	6,574 (7.36%)	693 (4.47%)
nai	6,489 (7.52%)	6,426 (7.29%)	1,318 (8.51%)
teinei	4,304 (4.99%)	4,815 (5.46%)	1,029 (6.64%)
suitei	2,029 (2.35%)	2,374 (2.69%)	250 (1.61%)
you	1,880 (2.18%)	1,799 (2.04%)	259 (1.67%)
meirei	1,570 (1.82%)	1,331 (1.51%)	517 (3.34%)
tekudasai	1,126 (1.30%)	994 (1.13%)	444 (2.87%)
tekuru	1,040 (1.21%)	1,068 (1.21%)	302 (1.95%)
joutaihenka	930 (1.08%)	1,417 (1.61%)	358 (2.31%)
tai	905 (1.05%)	895 (1.01%)	51 (0.33%)
tekureru	719 (0.83%)	707 (0.80%)	128 (0.83%)
teshimatta	639 (0.74%)	526 (0.60%)	58 (0.37%)
gimu	574 (0.67%)	648 (0.73%)	89 (0.57%)
dekiru	542 (0.63%)	813 (0.92%)	80 (0.52%)
sase	521 (0.60%)	694 (0.79%)	71 (0.46%)
darou	505 (0.59%)	534 (0.61%)	184 (1.19%)
teiku	399 (0.46%)	427 (0.48%)	57 (0.37%)
temiru	391 (0.45%)	399 (0.45%)	60 (0.39%)
teoku	326 (0.38%)	356 (0.40%)	59 (0.38%)
sugiru	326 (0.38%)	317 (0.36%)	44 (0.28%)
souda	291 (0.34%)	311 (0.35%)	63 (0.41%)
utosuru	249 (0.29%)	232 (0.26%)	30 (0.19%)
desu	215 (0.25%)	311 (0.35%)	28 (0.18%)
kaishi	193 (0.22%)	211 (0.24%)	84 (0.54%)
teshimau	189 (0.22%)	298 (0.34%)	75 (0.48%)
tearu	189 (0.22%)	205 (0.23%)	64 (0.41%)
yotei	65 (0.08%)	89 (0.10%)	4 (0.03%)
rashii	53 (0.06%)	72 (0.08%)	23 (0.15%)
nisuru	46 (0.05%)	198 (0.22%)	16 (0.10%)
teyaru	45 (0.05%)	61 (0.07%)	7 (0.05%)
teyoi	42 (0.05%)	40 (0.05%)	2 (0.01%)
noda	38 (0.04%)	38 (0.04%)	23 (0.15%)
hougayoi	31 (0.04%)	33 (0.04%)	2 (0.01%)
katei	3 (0.00%)	2 (0.00%)	1 (0.01%)
kirezu	1(0.00%)	0 (0.00%)	0 (0.00%)
合計	86,295	88,193	15,495

考慮した被覆率評価試験により，文型パターン数と被覆率の関係を調べる必要がある[12]．

(2) 文要素の任意化の効果

文型パターン化では，各種の変数と関数のほか任意化などのための記号類が使用される．このうち，任意化の機能によって縮退する文型パターン数を調べた．その結果によれ

[12] 文型パターン作成で使用した標本文を入力文として使用するクロスバリデーション法により，各入力文の文要素の何%が文型パターンで解釈可能かを考慮した被覆率実験では，単語レベル，句レベルの文型パターンの被覆率は，それぞれ 70 %，89 %，78 %であった．なお，詳細については，別途報告する．

ば，前者では,「パターン要素任意化」によって文型パターンが縮退する割合は 0.2%にとどまった[13].

5　あとがき

言語表現の構造と意味の関係に関する線形性と非線形性の定義に基づき，日英対訳例文から，非線形な表現構造を取り出して文型パターン化する方法を提案し，そのための具体的な作業手順を示した．また，その方法を重文，複文 15 万件の日英対訳例文に適用し，単語レベル，句レベル，節レベルの３種類の文型パターンをそれぞれ 12.8 万件，10.5 万件，1.3 万件（合計 24.6 万件，異なり 221.1 万件）を試作した．

単語レベルと句レベルの文型パターンでは，標本とした対訳文に含まれる自立語（平均６語）のうち３～４語が線形要素として変数化されたこと，また，関数化される付属語が１件弱あったことから，文型当たりの線形要素は平均４～５カ所存在することなどが分かった．節レベルの文型パターンでは，変数化された節は４～５%に過ぎず，大半の節は非線形要素であった．

これらのことから，重文，複文の構造は，殆どが非線形で，節に分解して翻訳する要素合成法は適さないが，単語や句のレベルで見るとかなり多くの線形要素を含んでいることが分かる．

また，人手による文型パターンの作成では，膨大な作業量が予想されたが，作業基準の明確化と各種のプログラムツールの準備などにより，作業の大半を自動化することができた．人手作業に頼る場合に比べて作業量は約 1/10 に削減したと推定され，作業品質の均一化も達成することができた．

従来，汎用的で大規模な文型パターン辞書の開発は困難と考えられ，対象分野を限定した用例翻訳などが試みられてきたが，上記のように標本文には多くの線形要素が存在すること，文型パターン化作業の大半が機械化できることから見て，汎用的な文型パターン辞書開発の展望が得られた．

ところで，今回作成した文型パターンは，文法情報で記述されている．試行的な実験によれば，再現率はかなり高い値が得られる見込みであるものの，意味的な適合率はまだまだ低い値だと予想される．意味的な適合率を向上させるには，各種変数に対する意味的な制約条件の付与が必須と見られる．今回の試作によって多くの文型パターンを得ることができ，種々の実験的検討が可能となった．今後は，(1) より高度な汎化手法とその効果の推定，(2) 意味的制約条件付与の粒度とその効果の推定などについての実験的検討を進め，被覆率と意味的排他性ともに優れた文型パターンを目指して改良を行っていく予定である．

[13] パターン数圧縮の効果は予想以上に小さい値となったが，この「パターン任意要素」の表現法は，得られた文型パターンの適用範囲を拡大することを意図したものであり，被覆率向上効果は期待できる．これについては別途実験的に確認する予定である．

謝辞

この研究は，科学技術振興機構（JST）の戦略的基礎研究事業（CREST）で行ったものである．ご議論を頂いた宮崎正弘氏（新潟大），池田尚志氏（岐阜大），新田義彦氏（日本大），佐良木昌氏（長崎純心大）山本和英氏（長岡技科大），白井諭氏（元 ATR），横尾昭男氏（NTT）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株式会社の皆様に感謝する．

参考文献

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，Vol.28, No.12, pp.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」，全５巻，岩波書店

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，Vol.16, No.3, pp.522-430

池原悟，佐良木昌，宮崎正弘，池田尚志，新田義彦，白井諭，柴田勝征 (2002)：等価的類推思考の原理による機械翻訳方式，電子情報通信学会，言語と思考研究会 TL2002-34, pp.7-12

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，思考と言語研究会 TL2002-48，pp.1-6

池原悟（2003）：言語で表現される概念と翻訳の原理，（TL 研究会 03-12 予定）

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（NL 研究会 04.1 予定)

池原悟，村本奈央，徳久雅人，村上仁一，宮崎正弘，佐良木昌 (2004)：機械翻訳のための日英文型パターン記述言語の設計 (NLP 投稿中)

衛藤純司，池原悟，池田尚志，佐良木昌，新田義彦，柴田勝征，宮崎正弘，白井諭 (2003)：意味類型構築のための接続表現の体系化について，情報処理学会研究報告，2003-NL-155，pp.31-38(2003.5/26/27 東工大)

金出地真人，徳久雅人，村上仁一，池原悟 (2003)：結合価文法による動詞と名詞の訳語選択能力の評価，情処研報，自然言語処理研究会，2003-NL-153-16，pp.119-124(2003.1)

長尾真 (1996)：「自然言語処理」岩波書店

長尾真，黒橋貞夫，佐藤理史，池原悟，中尾洋 (1998)：岩波講座「言語の科学」第９巻「言語情報処理」，岩波書店 (1998)

三浦つとむ (1967)：「言語と認識の理論」第１～３巻，勁草書房

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．

阿部　さつき： 1987 年東洋大学文学部国文科卒業．2000 年筑波大学修士課程経営・政策科学研究科経営システム科学専攻修士課程終了．1987 年（株）カンテックに入社．1992 年よりＮＴＴアドバンステクノロジ株式会社に入社，現在に至る．日英機械翻訳システムの辞書・ルール構築等の言語分析業務に従事．情報処理学会，言語処理学会，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．

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)